UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Computational study of pH-dependent membrane insertion mechanism of pHLIP peptides

Tomás Fernandes da Silva

MESTRADO EM BIOQUÍMICA
Especialização em Bioquímica

Dissertação orientada por:
Doutor Miguel Ângelo dos Santos Machuqueiro
Doutor Diogo Ruivo dos Santos Vila Viçosa

2017

Acknowledgments

In the first place, I would like to acknowledge my two supervisors: Miguel Machuqueiro and Diogo Vila Viçosa, for whom I am thankful for unveiling and having the patience to walk me, even if only a little, through the world of MM/MD simulations and, above all, for doing it without a single dull day. I would also like to acknowledge Bruno Victor, Paulo Costa and Rafael Nunes for the fun lunches, coffee breaks and help they provided me with. They were an always helping hand allied with a remarkable sense of humor. Finally, I am grateful to Pedro Reis for being an outstanding friend and for, one fateful day, having turned to me and say: "Hey! I think you will like this kind of stuff. Do you want to talk with Miguel? I have already texted him. "

I would also like to thank Ana Cruz, André Nascimento and Patrícia Rodrigues for being very close friends during these last few years. I would have made it through Bachelor's and Master's anyway, but it would have not been as fun as it was. Some people also deserve a very special mention: Pedro Monteiro, Daniel Tomás, Joaquim Veiga and Ana Rita Pratas thank you for the invaluable support and extremely long conversations.

On a more familiar note, I must appreciate all the effort made by my parents, Maria José and José Manuel Silva, and grandparents, Teresa and Dionísio Fernandes. Not only on the economical side as they supported me as I grew up and even now, but especially on the supportive and educational side. Now, I would not be half the person that I am and, especially, the one I try to be everyday, if not for them. My siblings Beatriz, Leonor and Dinis Silva for being just a bit too loud and always amusing me with their shenanigans.

Last and most certainly not least, a special acknowledgment to Beatriz Pimenta. Thank you for being the emotional support that I did not know that I needed and for being by my side, everyday, for the last few years. At the very least, you do not hate me anymore as much as you did the first time we met... I think.

Abstract

The pH (low) insertion peptide (pHLIP) belongs to a family of peptides originated from a segment of the transmembranar C helix of bacteriorhodopsin. The peptide has three major states that it may adopt when co-existing with lipidic membranes: state I - soluble and unstructured; state II - adsorbed at the membrane surface and unstructured; state III - inserted in the bilayer as an α-helix at low pH values. One of the major applications of pHLIP befalls on this ability to insert in membrane cells with an acidic vicinity, such as tumoral cells, thus working as an efficient tumor-specific biomarker. However, *wt*-pHLIP has a major limitation, since it also accumulates in the kidneys in considerable amounts due to their naturally acidic extracellular pH. There is a need to increase pHLIP specificity by delimiting the pH range of insertion to further improve its application as a biomarker and possible drug-delivery system for inflammatory tissues. At the beginning of this thesis, several calculations were already performed with a linear response approximation (LRA) method for *wt*-pHLIP and some variant sequences. Additionally, it was calculated the solvent exposure of a residue when inserted in the membrane, defining insertion dependent pK_a profiles. These profiles are crucial to understand the mechanism behind the withdrawal and insertion processes. However, there were some methodological limitations in LRA that propelled the need to use other sampling methodologies and a new L16H sequence to study the system. The main purpose of this thesis was to study pHLIP peptides using other sampling methodologies, namely constant-pH molecular dynamics (CpHMD) and pH-replica exchange (pHRE) methods, and evaluate the quality of the methods comparing with the LRA and experimental results.

The CpHMD method presented some limitations sampling high-energy protonation states, revealing an insufficient amount of data to extensively describe pK_a profiles. The newly developed pHRE method allows the exchange of pH values between replicas within a certain probability, thus facilitating transitions between energy minima and sampling of non-favorable protonation states, further enhancing the sampling of these states. Furthermore, the method to calculate insertion values was optimized. This new method is capable of defining local deformations and take them into account when calculating the insertion of the residues. Ultimately, these new methodologies were applied for simulations of *wt*-pHLIP and L16H variants. Several parameters were tested: number of lipids, different exchange frequency attempts and different pH ranges and steps.

In summary, the pHRE method presented some remarkable results by overcoming the sampling limitations of the previous methods, allowing more detailed, accurate and consistent pK_a profiles and pK_a^{ins} of residues. Likewise, the new insertion method presented a robust and reliable method to depict the gradient of solvent exposure of a residue in the membrane. Moreover, the L16H results corroborated the LRA results, considering that the L16H variant is able to delimit the range of pH insertion, despite the ΔpK_a^{ins} being too large for *in vivo* studies and therapeutic purposes. For the future, we now have the necessary tools to further expand the number of pHLIP sequences mutated with cationic residues, as to narrow down the pH range of insertion and evaluate the validity of these variants for clinical applications.

Keywords: pHLIP, pHRE, pK_a^{ins}, pK_a profile, distance cutoff method

Resumo

O pHLIP (*pH low insertion peptide*) pertence a uma família de péptidos derivados de um segmento da hélice transmembranar C da bacteriorodopsina. Este péptido consegue adoptar um de três possíveis estados: estado I - sem estrutura definida em solução de pH alto; estado II - adsorvido na superfície da membrana ainda sem estrutura definida; estado III - inserido ao longo da membrana adoptando uma hélice α na presença de valores baixos de pH. No caso do péptido pHLIP, o *folding* do mesmo sucede simultaneamente com o processo de inserção, sendo bastante dependente das interacções electrostáticas que estabelece entre os seus resíduos e a membrana. Estas características do pHLIP proporcionou um amplo estudo de possíveis aplicações terapêuticas, entre as quais como um biomarcador tumoral específico e como um transportador de fármacos para tecidos inflamados. O cerne destas aplicações prende-se com a especificidade do pHLIP para ambientes com valores de pH acídicos, principalmente células tumorais devido ao efeito de Warburg. Diferentes sequências deste péptido foram sintetizadas com o propósito de aumentar a especificidade do mesmo para as células tumorais, viabilizando métodos de tratamento e diagnóstico. No entanto, a especificidade do pHLIP é comprometida dado que acumula nos rins, devido ao pH naturalmente acídico. Desta forma, a acumulação de pHLIP nos rins tem sido abordada como uma limitação e um problema a resolver. Com este propósito, existe a necessidade de delimitar o intervalo de pH's de inserção, somente permitindo inserção a moléculas que se encontrem dentro de um intervalo específico de pH.

Ao iniciar esta tese, já tinham sido concretizadas diversas simulações do *wt*-pHLIP e de uma variante L16H utilizando o método de LRA (aproximação de resposta linear). Esta variante possui um resíduo catiónico na posição 16 e, segundo os resultados obtidos, confere dois pK_a^{ins}, desta forma delimitando os pH's nos quais pode inserir. Para obter estes resultados tinha sido utilizado um método de inserção média para avaliar o gradiente de exposição dos resíduos ao solvente, calculando qual o grau de inserção de cada um na membrana. Este método calcula a inserção de cada resíduo ao utilizar como referência a média das coordenadas Z dos grupos fosfato dos lípidos. Utilizando esta medida de inserção, é possível criar perfis de pK_a^{ins} ao longo da normal da membrana. Estes perfis permitem avaliar os processos de inserção e saída ao caracterizar o comportamento entre os resíduos chave do pHLIP. Não obstante, o LRA e o método de calcular inserção usados possuem algumas limitações que não permitem a melhor descrição tanto do sistema *wt* como da variante L16H. Deste modo, tornou-se imperativo utilizar diferentes métodos de amostragem (o método de dinâmica molecular a pH constante, CpHMD, e o *pH replica exchange*, pHRE) para estudar o pHLIP e as suas variantes.

Inicialmente, realizaram-se simulações de CpHMD do pHLIP *wt* e L16H para averiguar a qualidade do método, compararando com os resultados experimentais e os de LRA. Observou-se que o CpHMD também sofre de problemas de amostragem ao não conseguir amostrar estados ionizáveis dos resíduos em regiões de inserção próximas das caudas dos lípidos. Esta limitação impossibilita a previsão de curvas de titulação e, consequentemente, o cálculo de valores de pK_a. O pHRE permite a troca de valores de pH entre réplicas (simulações independentes), levando a que os estados

de protonação iniciais passem a ser estados de maior energia e, como tal, não favoráveis. Isto permite que estes sejam amostrados, mesmo que no decorrer da simulação os estados de protonação alterem para estados mais favoráveis, melhorando a amostragem a diferentes pH's, contribuindo assim para o cálculo de curvas de titulação. O desenvolvimento de um método de inserção alternativo, o método de *cutoff* de distância, permite que zonas locais de deformação sejam correctamente amostradas ao ter em conta a média de coordenadas Z dos fosfatos da deformação. Assim sendo, quaisquer fosfatos que afectem a real superfície da membrana são atenuados na média. Para testar e averiguar a qualidade dos novos métodos desenvolvidos realizaram-se diversos grupos de simulações recorrendo ao *wt*-pHLIP e à variante L16H, onde foi possível variar os seguintes parâmetros: número de lípidos; frequência de tentativa de troca ; diferentes intervalos de pH.

Na análise de resultados foi possível observar que o método de CpHMD apresenta falhas significativas na amostragem de estados ionizáveis de resíduos em regiões bastante inseridas. Isto deve-se ao método favorecer a amostragem de estados de protonação mais prováveis a um determinado pH, não garantindo que os estados menos prováveis sejam amostrados a diferentes pH's equivalentemente. Ainda assim, o método do CpHMD conseguiu prever com um elevado grau de confiança o valor do pK_a^{ins} do Asp14. Este valor é comparável ao valor de pK_{ins} experimental, o que nos leva a crer que o Asp14 adopta um papel importante na regulação dos processos de inserção e saída. Na sequência desta análise, aplicou-se o método de pHRE ao sistema de *wt*-pHLIP, estudando os diversos parâmetros referidos com o intuito de os optimizar para as simulações da variante L16H e futuras sequências. Observou-se através de medições de espessura da membrana que, ao utilizar uma membrana pré-equilibrada de 256 lípidos, existe uma região relativamente extensa não afectada pela presença do péptido. Deste modo, é possível reduzir o tamanho do sistema ao remover lípidos da membrana, levando a um aumento da velocidade de simulação, tornando o sistema mais eficiente. Com esse propósito, realizaram-se simulações com 128 lípidos, revelando-se comparáveis com as simulações de 256, sendo possível observar deformações locais semelhantes. Simultaneamente, estudou-se o impacto da utilização de diferentes valores de τ_{RE} e intervalos de pH na amostragem e probabilidades de troca. Constatou-se que a utilização de 20 ps e de 100 ps, em termos probabilísticos, é equivalente para o mesmo intervalo e salto de pH, pelo que o uso de 20 ps e um intervalo de pH's de 4.0 a 7.5 pode-se revelar mais vantajoso pelo maior número de trocas que poderá efectuar, melhorando assim a amostragem dos estados de protonação.

Deste modo, utilizaram-se os parâmetros aqui optimizados nas simulações de L16H com pHRE com o intuito de revalidar o efeito da histidina no processo de inserção do pHLIP. De forma semelhante ao LRA e CpHMD, foi possível observar um intervalo de pH onde ocorre a inserção do péptido pHLIP. Nesse intervalo, tanto o Asp14 como a His16 se encontram neutros ao inserir na membrana. Não obstante, o pHRE oferece uma melhor amostragem dos diferentes estados de protonação dos resíduos em regiões mais inseridas da membrana comparativamente ao CpHMD e LRA, resultando em perfis de pK_a mais detalhados. Aliado a estes resultados, o pHRE conseguiu prever pK_a^{ins} comparáveis, dentro da estimativa do erro, não só aos medidos com o CpHMD e LRA mas também aos resultados experimentais do Asp14 e His16. Relativamente aos resíduos aniónicos na região do C-terminal, observou-se que a sua inserção aparenta seguir uma ordem determinada pelos respectivos perfis de pK_a. Deste modo, os processos de inserção e saída demonstram ser definidos pelo equilíbrio de cargas no C-terminal e o estado de protonação do Asp14, no caso do *wt*-pHLIP, e adicionalmente da His16, caso se trate da variante L16H.

Em suma, ao conseguir estudar a variante L16H do pHLIP recorrendo ao método de pHRE, cumpriram-se os objectivos chave desta tese. Demonstraram-se as limitações inerentes do LRA e do CpHMD na amostragem neste sistema e optimizaram-se os parâmetros de simulação para o *wt*-pHLIP e possíveis variantes. Foi possível reproduzir o intervalo de pH's de inserção para a L16H em concordância com dados experimentais ainda que estes não permitam atingir o propósito

desejado de um uso terapêutico. Ainda assim, cimentaram-se as pedras basilares para futuros objectivos tais como: a calibração do método usando outras sequências testadas experimentalmente, revalidação de resultados obtidos com o método de LRA e, por fim, estudos de novas variantes do pHLIP usando diferentes resíduos catiónicos como a lisina e os seus derivados não-naturais de cadeia curta. Estes novos estudos traçam o caminho para, possivelmente, melhorar a eficiência na utilização do pHLIP como um método terapêutico no diagnóstico e tratamento de tumores.

Palavras-chave: pHLIP, pHRE, pK_a^{ins}, perfil de pK_a, método de *cutoff* de distância

Contents

List of Figures

List of Tables

List of Abbreviations

Asp	Aspartic acid
ATP	Adenosine triphosphate
CD	Circular dichroism
CE	Continuum electrostatics
CpHMD	Constant-pH molecular dynamics
C-ter	C-terminus
GRF	Generalized reaction field
Glu	Glutamic acid
His	Histidine
l-BFGS	limited memory - Broyden–Fletcher–Goldfarb–Shanno
LPBE	Linearized Poisson-Boltzmann equation
LRA	Linear response approximation
MC	Monte Carlo
MD	Molecular dynamics
MM	Molecular mechanics
NMR	Nuclear magnetic resonance
N-ter	N-terminus
NLLS	Non-linear least squares
PB	Poisson-Boltzmann
PBC	Periodic boundary conditions
PDB	Protein data bank
PEF	Potential energy function
pHLIP	pH (low) insertion peptide
POPC	1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine
RE	Replica-exchange
REMD	Replica-exchange molecular dynamics
RF	Reaction field
RMSD	Root mean square deviation
SE	Standard error
SPC	Simple point charge
TM	Transmembranar
T-REMD	Temperature replica-exchange molecular dynamics
Trp	Tryptophan
wt	wild type

Chapter 1

Introduction

1.1 pHLIP: a pH (low) insertion peptide

1.1.1 Peptide Structure and Properties

Folding consists on the capability of a peptide or protein to acquire one or more local spatial arrangements called secondary structures that will interact with each other and stabilize themselves with several weak interactions, such as hydrogen bonds [1]. The global three-dimensional structural stabilization of secondary structures confers possible interactions, through hydrogen or dissulfide bonds, between them leading to tertiary structures. Finally, the tertiary structures may be fully folded proteins or, in some cases, subunits. The three-dimensional arrangement of subunits in complexes constitutes a quaternary structure.

One of the most adopted and stable secondary structures is the α-helix. Helical structures are stabilized by retaining well defined torsional angles between the residues at periodic and repeated distances, allowing the peptide chain to fold as a helix [1]. A special type of helical proteins are the transmembranar proteins that can insert partially or fully into a lipid membrane. These transmembrane proteins are characterized by one or more hydrophobic region(s) with 20 aminoacids each, that interact directly with the acyl chains, and by flanking polar regions, which interact with the polar head-groups [2]. The insertion process of transmembrane proteins can be described in two stages for the sake of simplification: the independent folding of each helix while interacting with the membrane and, for more complex proteins, the interaction between the helices to constitute a more organized and stable structure. Furthermore, there are cases where other macromolecular structures provide an auxiliary support in the insertion process [1, 3]. In the case of monomeric peptides like the pH (low) insertion peptide (pHLIP), the insertion is coupled to its folding process which is heavily influenced by the interactions between the residues themselves and the membrane.

pHLIP belongs to a family of peptides derived from a segment of the transmembranar C helix of bacteriorhodopsin [4]. For the last few years, pHLIP has been extensively studied from structure to the kinetics involved in state transitions and the insertion/folding process. These experimental studies have shown that these peptides have the ability to insert across a lipid membrane at low pH values, while being soluble at higher pH values [4, 5].

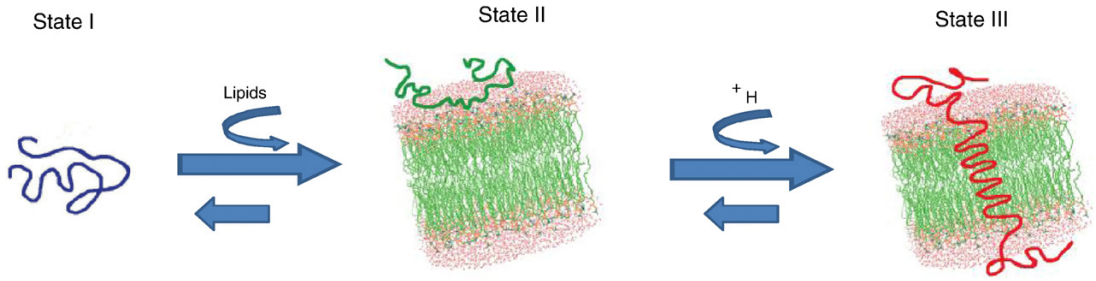

Figure 1.1: Scheme of the observed pHLIP states (adapted from [6])

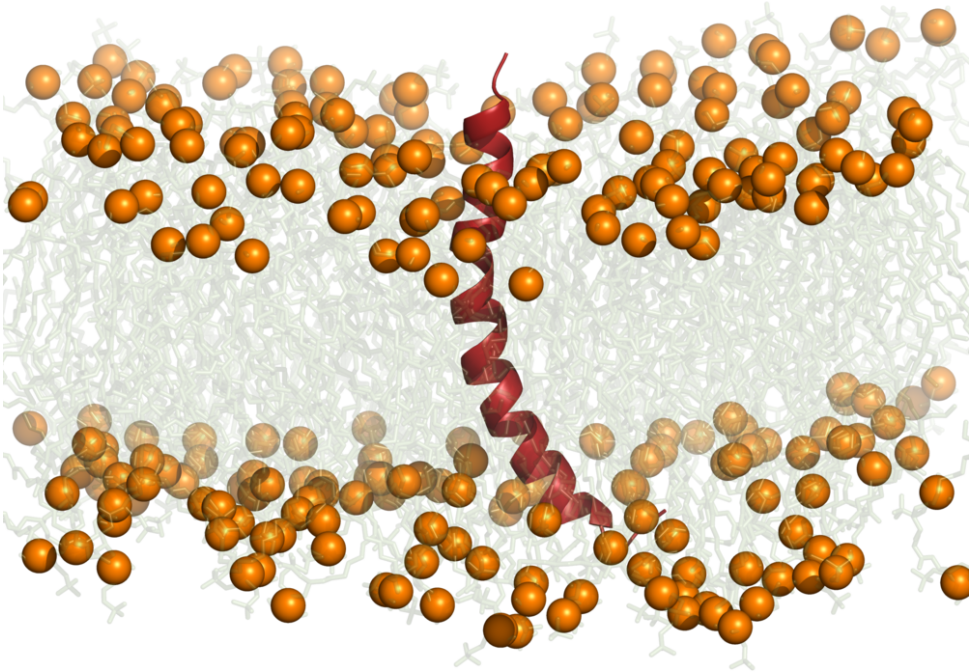

Figure 1.2: pHLIP peptide, represented in red cartoon, inserted across a POPC membrane bilayer obtained from a CpHMD simulation. Phosphour atoms are depicted as orange spheres, while the remaining lipid atoms are shown as transparent sticks for clarity.

Using experimental biophysical studies [4,5,7], it has been shown that pHLIP and its variants may adopt three possible states when co-existing with lipidic membranes (figure 1.1): in state I, pHLIP exists in solution without any visible secondary structures; in state II, pHLIP is weakly adsorbed to the lipid bilayer, remaining unstructured at a neutral pH (7.4); in state III, pHLIP is inserted across the lipid bilayer in an α-helix under acidic pH (5.0). This can be observed in changes of Trp fluorescence and circular dichroism (CD) spectra from experimental data which shows an increase of α-helical content as pH decreases [4, 8, 9].

With these features, pHLIP presents itself as a good case study for thermodynamic and kinetic studies of spontaneous membrane insertion and transmembranar protein folding and, simultaneously, to several practical applications.

1.1.2 Biological Relevance - Applications

Disease diagnosis has been a field of interest whose greatest concern resides in the speed and accuracy of methods, which led to new and promising applications. Although the study of tumors metabolic environment has been under scrutiny, making use of that knowldge has been a thorny challenge to surpass. In principle, we can target the specific acidic environment caused by the increased rate of anaerobic glycolysis observed in tumoral cells (Warburg effect) [10]. Due to hypoxia, this phenomenon leads to an increase of proton concentration in the cell. Although anaerobic glycolysis, per se, cancels out the proton [H+] production and consumption, the subsequent ATP breakdown allows the release of protons. Hence, it induces a cell response to pump out protons in order to preserve the cell pH homeostasis, and decreasing the extracellular pH to acidic values [11]. These values will be characteristic of tumoral cells and the metastization process.

As mentioned previously, pHLIP has the potential for several applications besides being an important model system [4, 5, 8, 9]. Previous works from Andreev et al. have shown that it is possible to take advantage of tumor acidity by attaching a small fluorophore at the N-terminus [4, 5, 8, 9]. Due to its ability to fold at low pH values in the presence of a lipid bilayer, pHLIP is able to fold and insert properly when present in the extracellular vicinity of tumoral cells, succeeding as a tumor-specific biomarker. Additionally, pHLIP is able to transport polar, cell-impermeable molecules through the bloodstream, since state I is water soluble, and translocate the same cargo (disulfide-linked to the C-terminus) across the cell membrane when it transitions from state I to state III [5]. Using tumor grafting on mice with subsequent pHLIP administration,whole-body fluorescent imaging [5] has shown that pHLIP is, indeed, able to target and accumulate in tumor cells with high efficiency. Nonetheless, whole-body and individual organ images have also shown that pHLIP accumulates in the kidneys in considerable amounts. This occurs since the kidneys possess an acidic extracellular pH and they are also an important organ for low-molecular-weight proteins catabolism, possessing several acidic regions which will promote pHLIP folding at those sites. As such, this effect presents a challenge to pHLIP specificity not only for acidic regions mapping but also for drug-delivery in inflammatory tissue associated with other diseases such as rheumatoid arthritis and atherosclerotic plaque formation [11, 12].

1.1.3 *wt*-pHLIP and Variants: Peptide Fine-Tuning

Peptide sequences can be fine-tuned through deletion, adition or replacement of key residues with the intent of changing intramolecular interactions, affecting their fold and global properties. Even if pHLIP has presented itself as a good model with relevant clinical purposes, there are some inherent flaws that need to be addressed [see section 1.1.2]. With that purpose in mind, several different variants have been synthesized for biophysical studies [9, 13]. The major purpose of these variants was to broaden the understanding of pHLIP interactions with the membrane, to obtain information on the specificity problem and to solve it by identifying and/or changing key residues of the wild type (*wt*) sequence. From the several variants that were synthesized, we should highlight the Glu and Asp variants (Table 1.1). The Asp variants are obtained by replacing the Glu with Asp residues in varying degrees of number of replaced residues, sequence order and truncation. The Glu variants are obtained with the same process with the replacement of Asp residues with Glu residues.

Table 1.1: pHLIP peptide sequence variants [9]. Underlined segments belong to the transmembrane region and bold residues are titrable key residues.

pHLIP Variant	**Sequence**
wt	ACEQNPIYWARYA**D**WLFTTPLLLL**D**LALLV**D**A**DE**GT
Asp	AC**DD**QNPWARYL**D**WLFPT**D**TLLL**D**L
	CDNNNPWRAYL**D**LLFPT**D**TLLL**D**W
	ACE**D**QNPWARYA**D**WLFPTTLLLL**D**
	ACE**D**QNPWARYA**D**LLFPTTLAW
Glu	AC**EE**QNPWARYL**E**WLFPT**E**TLLL**E**L
	C**EE**QQPWAQYL**E**LLFPT**E**TLLL**E**W
	C**EE**QQPWRAYL**E**LLFPT**E**TLLL**E**W
	AC**EE**QNPWARYA**E**WLFPTTLLLL**E**
	AC**EE**QNPWARYA**E**LLFPTTLAW

Comparing the emission and CD spectra at both neutral and acidic pH values, it was observed that the Glu variants have high er affinities for the membrane than their Asp counterparts, due to the additional methylene in the sidechain that contributes to a higher hydrophobic nature of Glu [9]. This information shows that Glu34 at the C-terminus of *wt*-pHLIP has an important role on membrane affinity in the transition from state III to state II. Also, it was possible to identify key residues for both folding and insertion processes (highlighted on *wt* sequence) such as Asp14 and Asp25, which are deep in the TM region (Table 1.1), as well as Asp31, Asp33 and Glu34 close to the C-terminus. [8, 14] Besides, there are also several more hydrophobic residues on pHLIP such as two Trp residues that were used to obtain a fluorescence signal.

As it was stated before, pHLIP will suffer state transitions due to the decrease in pH and membrane adsorption that will induce an overall structural change into a transmembranar helix. These two conditions will influence an important property and one of the main focus of this work: the pK of insertion.

1.1.4 Membrane Interaction with Key Residues - pK of Insertion

In the case of pHLIP, we know that pH and the pK_a of key residues have an important role in the insertion process of key residues. The relevance of those residues lies in: their ability to titrate and change pHLIP's solubility; its affinity to the membrane and the acidic side-chain interactions that strengthens or weakens the helix stability across the membrane. Thus, the solubility of pHLIP is regulated by these several titrable residues, more specifically, those that belong to the flanking sequence closer to the C-terminus which are Asp31, Asp33 and Glu34 [see Table 1.1]. These anionic residues will also regulate the rate of withdrawal of the peptide from the membrane bilayer since their protonation will start the process from the C-terminus [14]. Meanwhile, Asp14 and Asp25 are present in the TM region and they will have an essential role in this pH-dependent insertion mechanism since they need to be, at least, partially protonated so the peptide can remain inserted in the membrane [9, 14]. When these residues are protonated, they will be able to stay at either in the lipid tails region or at the ester/phosphate region without any electrostatic repulsion. Furthermore, a charged residue is stabilized when solvated by water due to the their inherent polarity and, only when these peptide residues are neutral, it can remain stable in that apolar

environment.

While using the fluorescence signal of the intrinsic Trp to follow the insertion process at different pH values, it was possible to determine a pK of insertion (pK_{ins}) [6, 8, 9]. Analogously to a pK_{a} value, pK_{ins} is the pH value at which half the population of pHLIP will be inserted. By evaluating the variants, it was measured a pK_{ins} of 4.5 for variants with a single Asp residue at the TM region. Meanwhile for Glu variants, it was shown an increase of approximately 0.5 units, which is close to the difference between those residues pK_{a} values [6]. This suggests that the apparent pK_{ins} depends on the protonation of at least one protonable residue and, indirectly, its intrinsic pK_{a} value. Furthermore, its pK_{a} value will be dependent on the dielectric of the membrane [6, 8, 9]. Hence for carboxyl groups, the pK_{a} will increase when the surrounding dielectric is low, which is the case of a membrane bilayer. Then, the more deeply imbued is the anionic residue, the more its pK_{a} will be shifted higher from their intrinsic value.

We can increase/decrease the population of protonated residues by manipulation of the pK_{ins}, which will allow or inhibit the insertion. Knowing that below a certain pH value, pHLIP will always insert itself (either in tumoral or kidney cells), we need to shorten that range. The addition of a cationic residue close to the aspartates may create an additional condition to the insertion process. This way, pHLIP will only insert into the membrane if both anionic (protonated) and cationic residues (deprotonated) are neutrally charged, otherwise it will remain in state II, adsorbed to the membrane. Following these lines, the variant L16H (sequence: ACEQNPIYWARYA**D**W**H**FTTPLLLL**D**LALLV**D**A**DE**GT) is here proposed as a test model. The possibility of creating a more specific pH range for the folding and insertion processes, by applying point-mutations on the *wt*-pHLIP sequence, may allow different tumoral cells to be targeted and, at the same time, to decrease the background noise in clinical imaging.

1.2 Computational Approach

Computational methods, such as molecular dynamics (MD), describe through statistical mechanics and the classical Newton laws of motion, a simulation of biomolecular processes and provide information on properties that are too short or too small to be accessible through experimental measurement [15, 16]. Another advantage is that any result may be presented through averages, distributions and time series of any desired quantity. Still, the model systems require experimental data to be reliably and accurately described. Lack of experimental results and innacurate data will make it impossible to describe the 'rules' that the system must abide by. In molecular mechanics (MM), these rules are known as the force field. A biomolecular force field consists on potential-energy terms that describe bonded (bond-stretching, bond-angle bending, improper and proper dihedrals) and non-bonded (Coulombic and Van der Waals terms) interactions between atoms [17]. These parameters will be fundamental descriptors of thermodynamic processes such as protein folding, protein receptor-molecule interactions, membrane formation and more. If we have energy, movement and time, through the classical laws of motion, a trajectory of a system of any size and in any time scale can be simulated with MD [15]. Simulations are still heavily limited by available computing power, however every year we have more powerful machines, and currently we can achieve the miliseconds timescale [18].

If a computational model can meet the requirements to correctly sample a biomolecular system, it will be capable of predicting properties that would be too difficult (if not impossible) or too expensive to determine through experimental work. Thus, using time and space averages, computer simulations are an important complement to experimental analysis by providing detailed information of complex biological systems with atomic detail.

A lot of biophysical studies have been conducted, leading to more collected and analyzed information on pHLIP, but our understanding of the insertion mechanism still remains pretty unclear. Although it is known which structural states pHLIP has and which residues have a crucial role on triggering insertion and allowing it to remain soluble, there's more detail yet to unveil. Tackling this problem with a computational approach will shed some light on important features of pHLIP such as pK_a values of the key residues, the influence of membrane insertion and the major contributors to the pK of insertion. This can prove highly advantageous since experimental studies cannot, at this time, accurately estimate specific residue contributions to a global pK of insertion nor sample and analyze, with atomic resolution, the insertion process.

1.2.1 The Sampling Problem: CpHMD and pHRE

Molecular systems simulations require that the sampled conformations of all the system configurations, which is called an ensemble, to be statistically representative [19, 20]. The number of degrees of freedom will have a direct impact on the number of configurations of the system. When the number of degrees of freedom is high, there will be a huge amount of configurations and correlations between them that will be dependent upon varying time and spatial scales [19,20]. The free energy surfaces described by the potential energy function will be rugged or smooth, with several or few basins depending on the amount of different configurations and energy states [fig.2].

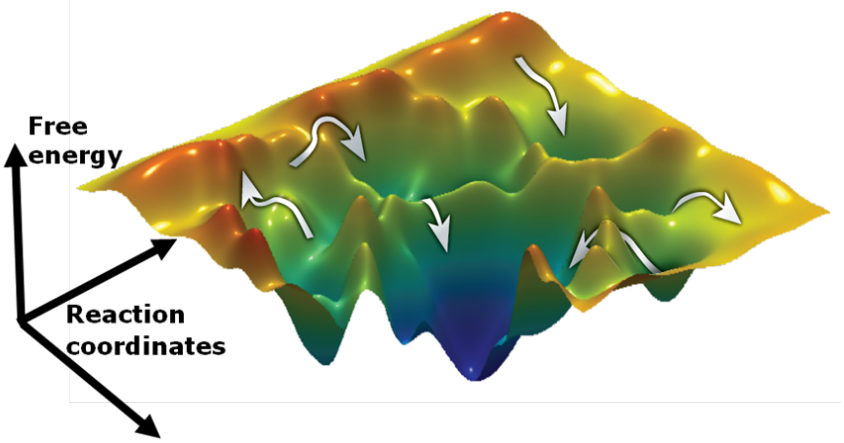

Figure 1.3: Representation of an energy surface for a given molecule adapted from [21]

Thus, sampling is closely related to the degrees of freedom of a system. To correctly evaluate properties of the system, statistically relevant results are needed. Simpler systems will have smaller phase spaces, the energy surface will be smoother and easier to search [22]. Whereas complex systems, like pHLIP, will be more diverse (higher degrees of freedom, rougher energy landscapes, more local energy minima) and less likely for all ensemble to be sufficiently sampled. The amount of time required to sample all the ensemble and its properties is proportional to: the complexity and size of the system; the desired degrees of freedom, either very detailed such as quantum mechanics (nuclei and electrons) or very approximate with atom groups (omission of degrees) both of which will influence the computational cost; the force-field accuracy and the search methodologies.

Several system parameters, like temperature and pressure, influence protein structure but there's a crucial property that has been troublesome to fully explore, and that is pH. pH effect on proteins has been shown to be extremely important since it will affect electrostatic interactions, one of the

strongest non-bonded forces at the molecular level. The pH effect can be substantially complex since it can define how the conformation populations will be distributed [23]. Regarding proteins, due to the possible numerous amount of titrable residues, each residue will have his unique local environment, affecting its pK_a, while several electrostatic interactions will be coupled between them, increasing the complexity.

In order to study the complex ways of pH and pH-induced protein structural changes, Baptista et al. [24,25] developed the stochastic titration constant-pH MD method (CpHMD) further expanded for peptides [23,26], proteins [27,28] and lipids [29–31]. The CpHMD method builds its basis on the complementarity between molecular mechanics/molecular dynamics (MM/MD) and continuum electostatics (CE). While MM/MD can simulate the dynamical behaviour of a system and sample the protein conformations [see section 2.1], it is not possible to take in consideration the protonable states of the residues as dynamic and interchanging throughout a simulation. On the other hand, CE methods, such as Poisson-Boltzmann (PB) calculations [see section 2.1.6 and 2.2] can describe the electrostatics of a molecule, though it can not explore the conformational space of the protein. Thus, CpHMD can sample proteins structures where pH affects their intramolecular interactions and, by extension, their conformations. This is accomplished by performing sequential blocks of PB/MC and MM/MD, where the former samples possible protonable states of the titrable residues, at a specified pH value, and the latter samples the conformational space.

In some systems such as membranes, sampling is still an issue to be adressed for CpHMD, MM/MD and other methods. Therefore, several studies have been conducted to improve sampling and different methodologies have been achieved for different purposes. Replica-exchange molecular dynamics (REMD) [32], metadynamics [33] and simulated annealing [34] are such methods with the purpose of diminishing, in different ways, the entrapment of the simulated systems [22]. When a protein, for example, is trapped on a low free energy minimum with a high energy barrier, it will take a long time before the protein is able to overcome it and, in that time, all the sampled information of the protein will be of that trapped conformational state.

Temperature replica exchange molecular dynamics (T-REMD), developed by Sugita and Okamoto in 1999 [32], carries out multiple independent parallel simulations at different temperatures. This method allows a periodic exchange of atoms positions between the RE simulations, depending on the temperature and probability of acceptance determined by a Metropolis criteria. Thus, this method possibilitates a mixed sampling in each simulation since a 'new' conformation will be introduced, pushing the simulation to different energy paths due to the higher temperature simulations allowing a faster sampling of different energy minima. In the end, new conformations will be sampled from the ensemble.

T-REMD is capable of improving sampling when using different temperature simulations, thus incrementing the CpHMD method with RE becomes a possible solution in further improving pH-dependent simulations. In the more recent years, these methodologies have been further expanded by Shen, Roitberg and others [35–38]. The pH replica exchange (pHRE) [39] establishes itself upon multiple CpHMD simulations with similar system parameters while using different pH values. Each CpHMD simulation will run using the same sequential blocks of PB/MC and MM/MD with an additional exchange step [39]. Analogously to the T-REMD, the concept of replica exchange can be extended to pH values of CpHMD simulations. In the end, each replicate trajectory will have, hopefully, exchanged several times with the others, thus improving the sampled conformations at each pH [see section 2.5].

1.3 Aim of this Work

The main purpose of this work is to study the pH effects on the conformational changes of the pHLIP peptide, both *wt* and its variant L16H. These pH effects have been studied using both the CpHMD and pHRE methodologies.

As previously mentioned, the insertion process and the pK of insertion are not thoroughly understood, neither if a cationic residue such as histidine is capable of delimiting the pH range for pHLIP to be clinically specific. Therefore, an evaluation will be made on the effects of replacing a residue on the pHLIP sequence with an histidine. It is crucial to understand how the addition of a potentially positively charged residue on an overall negatively charged enviroment influences the electrostatics of the key residues and, by consequence, the insertion process. This evaluation will be focused on pK_a predictions of key residues, whose environment is dependent on membrane insertion. The analysis of these results wil take into account previous results [40] and the experimental data from Andreev et al. [8].

Furthermore, there are several ways of measuring residue insertion. Since insertion directly impacts the environment surrounding a residue, it heavily influences pK_a predictions. Hence, a comparative study will be conducted between different insertion methods and how it reflects on the pK profile of each residue for the pHLIP variants (*wt* and L16H). In the end, it is also interesting to compare how well CpHMD and different pHRE setups fare against each other both in sampling the conformations of the pHLIP variants and predicting the residues pK_a values along the membrane normal.

Chapter 2

Theory and Methods

Following an introduction to the themes and ground work that composes this thesis, which were essential to lay the foundations to the work that was developed, this chapter will take an approach to the theories sustaining the used methodologies.

2.1 Molecular Mechanics/Molecular Dynamics

2.1.1 Molecular Mechanics

As mentioned in the Introduction (see section 1.2.1), as more detail exists in the system of interest or the higher number of degrees of freedom that are used to describe the system, more computing power is needed to accomplish the tasks at hand. Hence, when dealing with electronic motions, quantum mechanical methods are required to compute the interactions between the nuclei and the electrons. These studies are limited to smaller systems and larger ones require less degrees of freedom. Thus, using the Born-Oppenheimer approximation, one can remove the electronic motions and adopt the nuclear positions to calculate the energy of the system. With this simplification, the system is described by a molecular mechanics (MM) force field using a potential energy function (PEF) with parameters from an empirical force field. [19]

2.1.2 Potential Energy Function

To allow a valid description of the potential energy function of a system, each of the previously referenced components will be briefly explained. Every molecule, from the smaller ones to large polymers, obey certain types of intramolecular movements which can be mathematically described. These movements can be decomposed in bond stretches, angle bending and torsions. The following components presented are in a simple form. In a force field, more accurate forms may be used, though they require higher-order terms.

Regarding bond length stretching, the most elementary way to compute the movements between pairs of atoms is described by an harmonic potential based on Hooke's Law [19]:

$$\mathscr{V}(l) = \frac{k}{2}(l - l_0)^2 \tag{2.1}$$

Where l_0 is a reference bond length (the bond length where the energy is minimum), l is the length between the atom pair and k is the force constant between that specific pair. Although a simple approach, a true bond is not composed only of the stretch and the magnitude of the force. Complex interactions between several components will lead to adjustments on the energy which in turn leads to deviations from l_0. These energy variations can occur proportional to the square of the deviation from the reference bond length. Similarly to length stretching, angle bending also relies on an harmonic potential:

$$\mathscr{V}(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \tag{2.2}$$

Every angle is depicted as an energy deviation from its reference value (θ_0) dependent on a force constant (k). Even though bond stretches and angles are characterized by harmonic potentials, since the force constants are smaller as well as the reference values, the energy required to distort an angle is lower than to stretch or compress a bond. These molecular movements (bending and stretching) require substantial amounts of energy compared to the torsional or non-bonded terms. These two terms and how they interact with each other have a large impact on the structure. There are two type of torsional angles: proper and improper dihedrals. The proper dihedral is computed through a torsional angle formed by four sequential atoms (1-2-3-4) while improper torsional angles are defined by four non-sequential atoms (1-5-3-2). The proper torsional potential can be expressed as:

$$\mathscr{V}(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}[1 + cos(n\omega - \gamma)] \tag{2.3}$$

and the improper torsional potential can be expressed as:

$$\mathscr{V}(\omega) = k(1 - cos2\omega) \tag{2.4}$$

Where ω is the torsion angle formed between the pair of atoms that are 3 bonds apart. Note that the non-bonded interactions can also be computed with those same atoms comprising the torsion. The non-bonded interactions do not rely on any kind of specific bond interactions, though they are usually dependent on an inverse power of the distance. The non-bonded terms of a force field can be separated, at least, in two sub-components: the electrostatic and van der Waals interactions. The van der Waals interactions may rely on the Lennard-Jones potential, while the electrostatic interactions may be described using Coulomb's law.

An electrostatic interaction exists between two different molecules or between two regions of the same molecule where the only pre-requisite that exists is that they possess partial/net atomic charges. These charge distributions in molecules exist due to the varying electronegativity of the atoms in their constitution. As such, an electrostatic interaction can be calculated, with Coulomb's law, as a sum of pairwise interactions of point charges:

$$\mathscr{V}(r) = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \tag{2.5}$$

q_{i} is the partial atomic charges for atom i, N is the number of atoms in the system. r_{ij} is the distance between the atom pairs, while ε_0 is the permittivity in vacuum and ε_{r} is the relative dielectric constant of medium where the molecules are placed. The Van der Waals interactions account for the remaining non-bonded interactions in a system and compute both attractive (long

range) and repulsive (short range) forces between two atoms represented in the Lennard-Jones 12-6 potential function:

$$\mathscr{V}(r) = \sum_{i=1}^{N} \sum_{j>1}^{N} \frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}} \tag{2.6}$$

In the Lennard-Jones potential, there are the interaction parameters C_{12}, C_6 and the distance between the atoms (r_{ij}^{12} and r_{ij}^{6}). The interaction parameters refers to the repulsion (C_{ij}^{12}) and the attraction (C_{ij}^{6}) between atoms i and j. Finally, all the terms can be added to form the potential energy function for a system:

$$\begin{aligned}
\mathscr{V}(r^N) = & \sum_{bonds} \frac{k}{2}(l-l_0)^2 \\
& + \sum_{angles} \frac{k}{2}(\theta-\theta_0)^2 \\
& + \sum_{proper} \frac{V_n}{2}[1+cos(n\omega-\gamma)] \\
& + \sum_{improper} k(1-cos2\omega) \\
& + \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \left[\frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} + \left(\sum_{i=1}^{N} \sum_{j>1}^{N} \frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}} \right) \right]
\end{aligned} \tag{2.7}$$

The system description can be more or less accurate with such potential energy function $\mathscr{V}(r^N)$ and the level of detail will depend if higher-order terms are included or not. A system of N atoms with varying masses and Cartesian vector positions allows, in principle, an energy depiction of any configuration of the system. Now that we can depict the system configurations through the potential energy function, molecular dynamics will show how those configurations change over time.

2.1.3 Force Field

A force field contains several components and parameters in order to describe a molecular system, though every force field will eventually depend on four different components: bonds, angles, torsions and non-bonded interactions. Each component will contribute with an energy term to the potential energy function which will allow an interpretation of the entire system. In this work, the force field that was used was GROMOS 54A7.

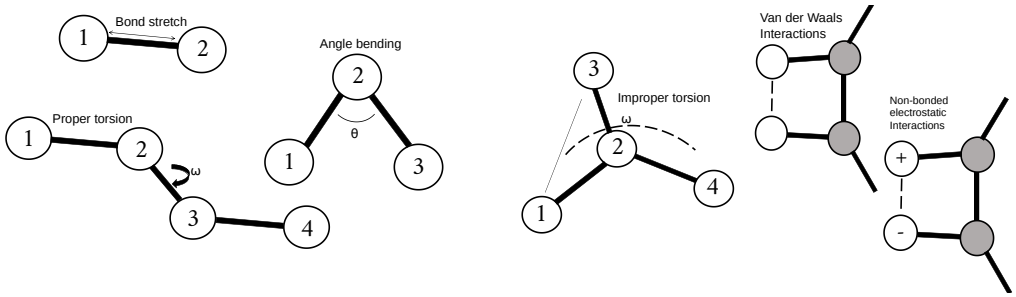

Figure 2.1: Schematic representation of the components that constitute a molecular force field: bonds, angles, torsions and non-bonded interactions (Van der Waals and electrostatic).

The GROMOS force field is comprised of specific parameters for atoms, bond, angles and dihedral types. Additionally, the GROMOS force field takes an united-atom approach which means that the apolar hydrogen atoms that are usually covalently bonded to aliphatic carbon atoms are not explicitly computed. Being colapsed into the carbon to which they are bonded to. The polar hydrogens, which play an important role in hydrogen bonding and other interactions are treated explicitly. Note that variability in these parameters will lead to different results even if the same functional forms are being used. Therefore, each force field is specifically parametrized using small molecules and further extrapolated to more complex systems. Different force fields must be used according to their purpose and how they were parametrized. [19, 20]

The diversity of molecular force fields derives from the fact that they are empirically designed, sustaining the idea that there is not one absolute force field to follow. Experimental data of small molecules fleshes out the force field, but the components of the functional form are the ones that give it shape. The four referenced parameters are very generic and it does not mean that they are all that is necessary to accurately describe the system. Neither does it mean that several more components in a functional form may provide high computational efficiency. Both extremes may lead to loss of information.

2.1.4 Molecular Dynamics

Molecular dynamics simulations study the movements and interactions of particles by integrating their coordinates with a descriptor of the potential energy using a molecular force field. As such, it is possible to calculate time averages of properties of any system. Since every particle in a system has a given mass, a trajectory can be simulated where the positions, forces and velocities will vary with time.

$$F_i = -\nabla_{ri} \mathscr{V} \tag{2.8}$$

F_i refers to the force acting on atom i derived from the gradient of potential energy (∇). The gradient will transform a scalar function such as the potential energy function into a vector function with individual differentials to each coordinate. This is useful to determine the steepest slope of a curve (like in an energy surface) which will provide an acceleration for atom i in any given point. That acceleration will be integrated for the determination of the new positions and velocities used to describe the trajectory of an atom in a MD simulation. Due to the high number of interacting particles, MD simulations of large systems become chaotic in its nature.

There are different methods to accomplish the integration steps of a MD simulation but one of the most used are the finite difference methods. These methods rely on two assumptions: the integration step is very small, allowing step-by-step calculations; the force is constant during the

time step. If we have a configuration at time t, every particle will have a force f associated to it. Then we can evaluate the positions and velocities of every particle at a time $t + \delta t$, obtaining new configurations and new velocities. The forces will then be recalculated for the next time step. Though, the larger the time step (δt) the less accurate the system will be described. Predicting a property is directly correlated to the degree of accuracy of the atoms positions. Which, in turn, are dependent on the frequency of update of the forces acting upon them. Higher frequency updates lead to higher accuracy on atoms position predictions, but also a higher computational cost.

A widely used type of algorithm that allows these calculations is the Verlet algorithm. One of its variants, the leap-frog, was used for this work. To compute the leap-frog algorithm, we need sets of positions ($r(t)$ and $r(t + \delta t)$) and accelerations $a(t)$. Besides that, the leap-frog algorithm follows these relationships:

$$r(t + \delta t) = r(t) + \delta t v(t + \frac{1}{2}\delta t) \tag{2.9}$$

$$v(t + \frac{1}{2}\delta t) = v(t - \frac{1}{2}\delta t) + \delta t a(t) \tag{2.10}$$

The leap-frog algorithm implements new velocities after a jump to a new time step $(t + 1/2\delta t)$. The velocities are not entirely necessary to predict the potential energy as perceived by eq. 2.8, though they are required to derive the kinetic energy and, consequently, the new positions of the system. Thus, the equations are necessary in order to determine the first trajectory velocities for the particles and to compute the following time steps positions and accelerations. The velocities at time t can be calculated as:

$$v(t) = \frac{1}{2}\left[v(t + \frac{1}{2}\delta t) + v(t - \frac{1}{2}\delta t)\right] \tag{2.11}$$

Periodic Boundary Conditions

Having successfully determined the configurations of the system overtime using the leap-frog algorithm, we are able to perform a MD simulation. We are capable of predicting certain properties and compute, at least, the most essential interactions between the molecules. However, in a simulation, the system is simulated in a box with finite size. The boundaries of the box must be treated to avoid wall effects that may create irregularities in the simulation. Periodic boundary conditions (PBC) enable simulations to be run with a small set of particles as if they were surrounded by a near-infinite amount of solvent molecules. A periodic system is characterized with the system of interest being simulated in the center while surrounded with images of itself.

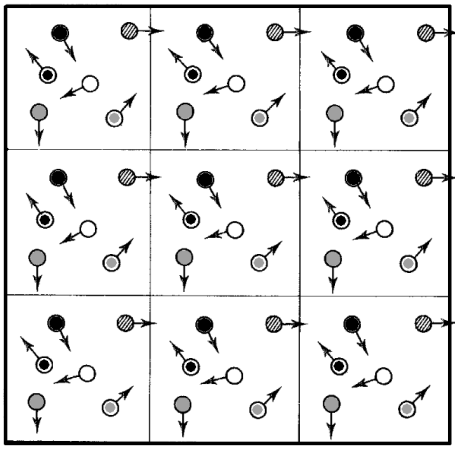

Figure 2.2: Graphical representation of periodic boundary conditions applied to a system in two dimensions (adapted from [19])

When trajectories are simulated, sometimes, a molecule may get closer to the boundaries of the box and eventually cross them (figure 2.2). When using PBC, every image of the particles will follow what happens in the original box, though multiplied by a proper vector depending on their coordinates. The particles will leave the box and be replaced on the other side of the box by an adjacent image. This allows consistency in the number of atoms in the box. Furthermore, the size of the box is important since any atom must not simultaneously interact with another atom and their image. Any error in the treatment of PBC may cause distorting effects in the simulation. In some cases, due to the nature of the method, the simulation may suffer from periodicity effects such as mimicking a crystal environment.

Treatment of non-bonded interactions

When considering small systems without PBC, the overall non-bonded interactions might be fully computed. Otherwise, larger and more complex systems using PBC, the number of non-bonded interactions will be infinite. The time spent on calculating the interactions for all atom pairs would be prohibitive. To avoid the excessive computational cost a limit must be set on the lists of atoms whose non-bonded interactions will be determined. For that purpose, it is common to use a cutoff method for the van der Waals interactions, while for the electrostatic interactions, the cutoff method can be used with the reaction field (RF) approach. The twin-range method can also be used to increase simulations speed by applying restrictions on the list of atoms that are 'worth' considering. The restrictions are placed in two ways: the use of cutoffs and the frequency of updates.

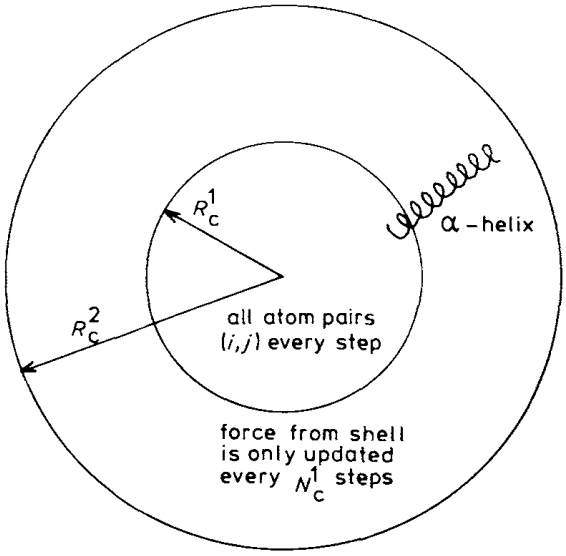

Figure 2.3: Schematic representation of the twin range method. R_C^1 and R_C^2 are the cutoff radii around particle i delimiting the number of atoms j interacting with it (adapted from [15]).

The first restriction relies on placing cutoffs around a particle i, thus delimiting the amount of atoms j that interact with i. In the twin-range method, there are two cutoff radii around particle i and this is where the second restriction comes into place. Every neighboring atom j, lying within the distance R_{short} (first restriction) will be considered and stored. For every N steps, the list of atoms whose distance to atom i is smaller than R_{short} will be updated (second restriction). When the distance to atom i is between R_{short} and R_{long}, every atom will be considered and updated but only every M steps (where $M > N$). Atoms whose distance to atom i is greater than R_{long} will not be taken into account.

The generalized reaction field (GRF) method is a particular case of the RF [41]. This method is used to simulate the effect of long range interactions in simulations that used PBC. In the reaction field method, some approximations are made to the system. First, if we take a molecule as an example, it is divided in two parts: an inner and an outer region. In the inner region, the atomic charges q_{i} are explicitly treated with a dielectric constant ε_1 as previously shown [see section 2.1.3]. In the outer region, instead of atomic charges, it is assumed that a continuous medium is surrounding the inner region with a dielectric constant ε_2 and an ionic strength I. The potential in the inner region is described by a direct Coulombic term that takes the charges q_i into account and a reaction field potential that is dependent on ε_2 and I. Thus, the electrostatic potential will be dependent on the outer region continuum since the medium will affect the interactions in the inner region. The outer region dielectric ε_2 and I ionic strength possess a shielding effect on the electrostatic interactions of the q_i particles. Therefore, the energy of the electrostatic interaction of particles q_i will be lower, the higher ε_2 is, since the interaction between the particles in the inner region is dampened. The effect of the ionic strength on the reaction field term is the inverse of the dielectric ε_2. For lower values of ionic strength and ε_2, an higher contribution will be given to the first Coulombic term. If $I = 0$, then the electrostatic potential is only dependent on the first term. The Poisson-Boltzmann equation becomes the Poisson equation, thus describing a RF instead of the GRF approach.

2.1.5 Pressure / Temperature

There are several parameters which will define a MD simulation. The phase space of a system is defined by coordinates and momenta and to specify the sampled ensemble during the simulation, some factors must be defined. Hence, temperature and pressure are necessary to be accounted for with the purpose of mimicking physiological conditions [20]. Some algorithms must be used to achieve these conditions. In a general note, the algorithms adjust the temperature by tweaking the velocities applied in each atom and the pressure by changing the positions of the atoms adjusting the system volume. In this work, the velocity-rescale (v-rescale) thermostat [42] and the Parrinelo-Rahman barostat [43] were used for temperature and pressure coupling, respectively, to sample the NPT ensemble [44].

Temperature

Most biomolecular systems in nature are under a constant temperature or small temperature range. Thus, controlling the temperature is essential during MD simulations considering that it will heavily influence the ensemble to be explored. In this work, the temperature is kept constant in an isothermal-isobaric ensemble (NPT) using the v-rescale thermostat which was originated from the Berendsen thermostat [45]. The temperature of a system is given by the average kinetic energy over time [44]. By coupling the system to an external bath, which either increments or removes energy through random velocity reassignment from the Maxwell-Boltzmann distribution [19, 44], the temperature will be fluctuating over the desired temperature T_0. Thus, system temperature deviations from T_0 will be corrected over time:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_T} \tag{2.12}$$

This thermal coupling will lead to a time-exponential decay of the temperature deviation (λ). The time and the strength of the coupling (τ_T) can be tweaked, where a small τ_T means a tight coupling and an high τ_T means a weaker one. Different values of τ_T allows different uses when preparing the system before a MD production. Smaller τ_T will be used for faster temperature decay and convergence to the desired T_0:

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left\{\frac{T_0}{T(t - \frac{\Delta t}{2})}\right\}\right]^{\frac{1}{2}} \tag{2.13}$$

Using this method, the fluctuations in the kinetic energy are not the same as the fluctuations that would exist in a canonical ensemble. Therefore, for small systems, this would lead to incorrect sampling. The v-rescale thermostat is essentially a Berendsen thermostat with a correction to the kinetic energy distribution:

$$dK = \left(K_0 - K\right)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_0}{N_{df}}}\frac{dW}{\sqrt{\tau T}} \tag{2.14}$$

An additional stochastic term was added dW, where K is the kinetic energy and dW is a Wiener process. The Wiener process gives time-dependent small random increments which will limit the rescale process of the kinetic energy. This stochastic term will bring a more correct depiction of the ensemble. The temperature coupling for the solvent (water) and the solute (pHLIP and

Membrane) was made separately to avoid phenomena of solvent and solute being in different temperatures while the overall system temperature is constant at the desired value.

Pressure

Analogously to the temperature, the isothermal-isobaric ensemble will also require a constant pressure. The Parrinello-Rahman method enables more realistic results to compare with experimental measurements by coupling the solvent and the solute to an external pressure bath. The fluctuations in pressure are more frequent and greater than those in most other factors.

The pressure is closely related to the atoms positions and the simulation cell volume. In the NPT ensemble, the pressure remains constant by creating fluctuations in the volume of the cell until convergence is achieved. The frequency of volume fluctuation is directly related with the isothermal compressibility (κ):

$$\kappa = -\frac{1}{V}\left(\frac{\delta V}{\delta P}\right)_T \tag{2.15}$$

The method used for pressure adjustment is analogous to the v-rescale thermostat, since it was used a pressure bath [45].

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}\Big(P_{bath} - P(t)\Big) \tag{2.16}$$

where τ_{P} is the coupling constant (an higher τ_{P} corresponds to smaller pressure coupling), P_{bath} is the desired pressure of the bath and $P(t)$ is the pressure at any given time t.

The τ_P is a measurement of the relaxation time of the system between volume variations of the simulation box. Thus, the relaxation time is associated with the frequency of the volume variations. At the same time, the system must also rescale the atoms velocities. When there is a high τ_P, the system will have more time to adjust to the different volumes. Hence, the system has a smaller pressure coupling, which is the same as needing more time until it achieves the established pressure P_{bath}. A smaller τ_P translates into a higher pressure coupling (more frequent volume variations) and the convergence of P to P_{bath} is faster. Nevertheless, systems with a low τ_P may not adjust fast enough to the volume variations, leading to undesired effects. Meanwhile, systems with a high τ_P will achieve the desired P slowly, but in a steadier way.

Energy Minimization

The basis of these computer simulations are defined by the potential energy function and a small set of parameters (like the pressure, temperature and number of particles) that characterizes the thermodynamic state of the system. Additionally, the Born-Oppenheimer approximation allows the system to be described entirely by the motion of the nucleus, because it assumes that the fast electronic motion will adjust to the slower nuclear motion thus decoupling both. As a consequence, each state of the system is described with coordinates of a Cartesian system composed of N atoms. The phase space is a multidimensional space that comprises the different combinations of mechanical states that the system can assume with a specific potential energy ($V(r_{\mathrm{n}})$) at a certain defined set of values for the parameters (pressure, temperature and number of atoms), in the case of a NPT simulation.

Certain system states will possess lower energies than others and the more complex a system is the more diverse the system states will be. An energy minimum will be a stable state of the system although not always the most populated one. Therefore, in a computer simulation, it is generally a good idea to lead the system to more stable states and then search and sample the conformational space of the system as best as possible.

The steepest descent method is a robust and easy to implement first-order minimization method. This method performs a downhill search on the phase space until it hits the closest energy minimum (not necessarily the global energy minimum). The method defines a vector r as a vector of all 3N coordinates and an initial maximum displacement. At $s = s_0 + 1$, where s is a step, new positions are calculated by:

$$r_{n+1} = r_n + \frac{F_n}{max(|F_n|)} h_n \tag{2.17}$$

where h_n is the maximum displacement of the atoms relative to the previous structure, F_n is the force applied in the atom components and $max(F_\mathrm{n})$ is the absolute maximum value of the force components. After the potential energy and the new positions are calculated for the new step, it is compared with the previous energy values. If ($\mathscr{V}_{\mathrm{n+1}} < \mathscr{V}_\mathrm{n}$), the new positions are accepted and the maximum displacement value is increased, leading to an higher jump in positions in the next step. If ($\mathscr{V}_{\mathrm{n+1}} > \mathscr{V}_\mathrm{n}$), the new positions are rejected and the maximum displacement value is lower than the previous one, leading to a smaller jump in the search. As the new positions get closer to the local minimum, the values of displacement will be lower. The search will stop if a specified number of steps as been performed or if a $max(|F_\mathrm{n}|)$ threshold as been crossed, possibly indicating a local minimum.

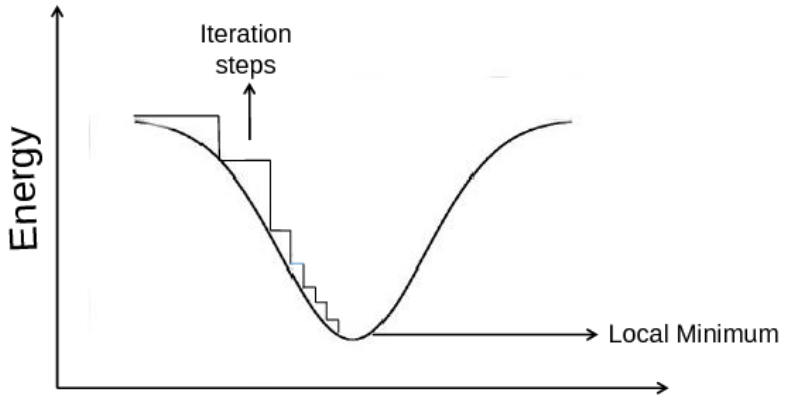

Figure 2.4: Graphical representation of a steepest descent method iterating over different energy values until it converges to a local minimum.

To complement the surface search of the system, other minimization methods are used after an initial use of steepest descent. Indeed, the limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newtonian minimizer [46], more easily remembered as the l-BFGS algorithm, may be used after a steepest descent. The main advantage of the method concerns the ability to search other local minima by performing jumps between them. On the other hand, this method is considerably slower than the steepest descent, more so if it does not start from a local minimum. Hence, this algorithm is usually performed as a second minimization step.

2.2 Continuum Electrostatics

The electrostatic potential is crucial in several biochemical processes, since the interaction between charged residues and polar groups will define protein affinity to solutes, intra-molecular interactions defining fold properties and inter-molecular interactions like protein binding, macromolecular complex formation and enzymatic reactions. In a biomolecular system, for example a protein in water, all the system atoms may be explicitly computed including both the protein and the solvent. Despite being possible to compute all those interactions, it would be impractical and computationally expensive.

In continuum electrostatics (CE), a common assumption is that the solvent is implicitly contributing to the non-bonded interactions through a medium with a high dielectric constant (ε) and an ionic strength I. Meanwhile, the protein will have a lower dielectric constant due to the several dipoles along its structure that remain constrained. A medium with high dielectric constant such as water (usually 80) will dampen the electrostatic interactions, creating a shielding effect between point charges. A CE model is composed of a low dielectric region (the protein) immersed in a medium with a high dielectric constant, whose separation is attained by a solvent accesible surfece and a ion exclusion boundary.

Assuming these conditions, an electrostatic potential can be calculated as a continuous function in space. For the non-bonded interactions, we resort to the Poisson-Boltzmann (PB) equation for a rigid body, which is an approximation of the system model.

$$\nabla.\varepsilon(\vec{r})\nabla\phi(\vec{r}) - \kappa^2 sinh\big[\phi(\vec{r})\big] = -4\pi\rho(\vec{r}) \tag{2.18}$$

Further derivation of the equation, by reducing the Taylor series to the first term only, allows the treatment of the interactions to be performed by the linearized PB equation (LPBE):

$$\nabla.\varepsilon(\vec{r})\nabla\phi(\vec{r}) - \kappa^2\varepsilon(\vec{r})\phi(\vec{r}) = -4\pi\rho(\vec{r}) \tag{2.19}$$

Decomposing the LPBE is essential to understand what lies beneath. First and foremost, this equation gives the electrostatic component for free energy calculations, thus it is dependent on the dielectric constant of the medium ε and how it varies with the relative position of point charges, hence $\varepsilon(\vec{r})$ where $\vec{r}$ is a position of a point charge in the system. Therefore, having characterized a system by the dielectric constant ε and the distribution of the point charges throughout the system (charge density $\rho(\vec{r})$), the electrostatic potential $\phi(\vec{r})$ can be determined as a solution of the Poisson-Boltzmann equation. Nevertheless, the electrostatic potential $\phi(\vec{r})$ also depends on the ionic strength I. Since the ions present in the solvent possess motion that is dependent on the charge distribution, their effect can be described with a term derived from the Debye-Huckel theory:

$$\kappa(\vec{r}) \begin{cases} \left(\frac{8\pi e^2 I}{\varepsilon_{out} K_B T}\right)^{1/2}, \text{if } \vec{r} \text{ is an accessible region to ions} \\ 0 \end{cases} \tag{2.20}$$

where $\kappa(\vec{r})$ is the Debye length that dictates the limit distance from which any charge q_{i} can be shielded due to the ion effect, $\varepsilon_{\mathrm{out}}$ is the solvent dielectric, k_{B} is the Boltzmann constant and T is the temperature. For regions where the ions are inaccessible, such as the hydrophobic regions of a protein, the Debye length is 0, thus the equation translates on the previous equation (eq. 2.20). If the ions do not have access to those regions, then the equation takes the form of the Poisson equation:

$$\nabla.\Big[\varepsilon(\vec{r})\nabla\phi(\vec{r})\Big] = -4\pi\rho(r) \tag{2.21}$$

Since analytical solutions of these equations do not exist for complex geometries, the solutions can be obtained using numerical methods, such as the finite difference methods. The finite difference method applies all the physical quantities described in the equation on a cubic grid with a certain grid space, promoting iterative calculations on each grid point. After a focusing step, these calculations will continue until they achieve a convergence criterion.

The electrostatic potential allows us to calculate electrostatic energies (W) of different states (eq. 2.22). These electrostatic energies can describe protonation states using the electrostatic interactions between the residues. Therefore, calculating the difference between electrostatic energies of two different protonation states (ΔW) allows a prediction of pK_a values(ΔG).

$$W = \frac{1}{2}\sum_i q_i\phi(\vec{r}_i) \tag{2.22}$$

2.2.1 Protonation Free Energy Calculations

The electrostatic energy gives us the ability to calculate the free energies of protonation of a titrable residue in a protein $\Delta G_{A\to AH}$. Since it is not possible to do it in a straightforward manner, a thermodynamic cycle must be used (Figure 2.5).

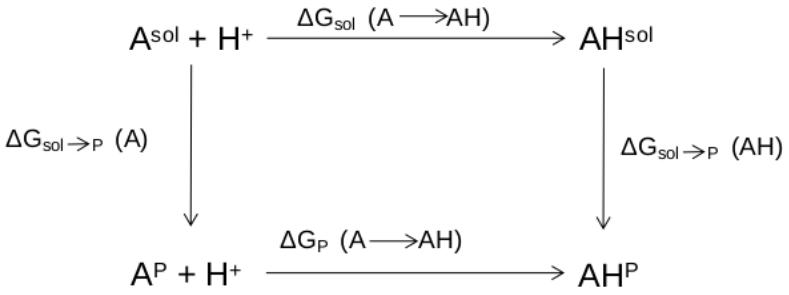

Figure 2.5: Thermodynamic cycle involving protein and a given model compound.

The free energy associated with the protonation of a titrable site in a protein, as described by a thermodynamic cycle, can be given by the terms that define a protonation event in a solvent environment and the respective transitions of each protonated state from a solvent to a protein environment:

$$\Delta G_P(A\to AH) = \Delta G_{sol}(A\to AH) + \Delta G_{sol\to P}(AH) - \Delta G_{sol\to P}(A) \tag{2.23}$$
$$= \Delta G_{sol}(A\to AH) + \Delta\Delta G_{sol\to P}(A\to AH) \tag{2.24}$$

Consider the following example, where A_{sol} and AH_{sol} refers to the deprotonated and protonated forms of any titrable site, respectively. Additionally, $A_{protein}$ and $AH_{protein}$ are the deprotonated and protonated forms of the same titrable site surrounded by the protein environment. The protonation free energy of $A_{sol}\to AH_{sol}$ can be translated into a pK_a value for the residue when surrounded by solvent molecules. Since we want to characterize the pK_a values of residues of interest in proteins,

we also need to use the free energies of $A_{sol} \rightarrow A_{protein}$ and $AH_{sol} \rightarrow AH_{protein}$ to complete the thermodynamic cycle. Consequently, it is deduced the $\Delta G_{protein}(A \rightarrow AH)$ and then converted into a pK_a value. As a matter of fact, it was used a different method regarding the pK_a of residue in a solvent. The pK^{mod} is an estimated value that is used to calibrate the pK_a of a simulated model compound to correctly match with the experimental pK_a of a residue. Therefore, any single residue pK_a value can be given by:

$$pK^{int} = pK^{mod} + \frac{1}{2.3\kappa_B T}\Delta G_{background}(P) \tag{2.25}$$

The pK^{int} will be calculated using the pK^{mod} while accounting for the contributions of the other residues in the protein that are not titrating $\Delta G_{background}(P)$ to obtain the final pK_a values of any residue. In a protein, we need to compute the contribution from the interaction between sites (W_{ij}).

$$pK_a = pK^{int} + W_{ij} \tag{2.26}$$

With the purpose of determining the probability of a certain protonation state $\vec{a}$, denoted as a vector $\vec{a}(a_1, a_2, a_i)$ where each term of the vector relates to a titrable site of the protein, we can calculate the energy difference between protonation state a_i and a reference state a_0 as follows:

$$\Delta G(pH)_{\vec{a_0}\rightarrow\vec{a_i}} = -2.3\kappa_B T \sum \vec{a}_i \gamma_i pK_i^{int} + \sum_i \sum_{j\neq i} \vec{a}_i \vec{a}_j \Delta\mathscr{V}_{ij} \tag{2.27}$$

where γ_i is the charge of the titrable site when it is ionized and $\Delta\mathscr{V}_{ij}$ is the interaction free energy between sites i and j. In the end, we have an equation for the protonation free energy with terms that can be obtained from PB calculations.

2.3 The Monte Carlo Sampling Method

Developing the idea of the previous section, the protonation state of any residue is dependent on the environment that surrounds it. Thus, if there is a large number of titrable sites in a protein, the neighboring residues and environment of each residue will dictate the protonation state which will further influence the electrostatic environment of the other residues. This complexity and diverse number of combinations of states for each conformation is present in macromolecular systems and it can not be easily nor quickly calculated. The Monte Carlo (MC) sampling method [47] is a technique that allows the generation of configurations of a system by allowing random movements in a simulation. By assigning random movements, we can explore the phase space of a system and deduce thermodynamical properties. Even though we can sample several configurations, the sampling will not be completely random. The selected configurations fall within a distribution of positions in a region of the ensemble that provide important contributions to the integration of the thermodynamic properties. It is possible to sample protonation states using a Monte Carlo method [24,47]. Since a transition between an ionized state $\vec{a}_i$ to a state $\vec{a}_j$ of a given conformation may be expressed as a $\Delta\Delta G$, it is similar to a transition between two conformations:

$$\Delta\Delta G_{\vec{a_i}\rightarrow\vec{a_j}}(pH) = \Delta G_{\vec{a_0}\rightarrow\vec{a_j}}(pH) - \Delta G_{\vec{a_0}\rightarrow\vec{a_i}}(pH) \tag{2.28}$$

Furthermore, it is also possible to compute the probabilities of certain protonation state at each pH as given by:

$$p(\vec{a}) = \frac{e^{\beta\Delta G(\vec{a}_i) - 2.3Z(\vec{a}_i)pH}}{\sum_{\vec{a}} e^{-\beta\Delta G(\vec{a}) - 2.3Z(\vec{a})pH}} \tag{2.29}$$

where $\beta = \frac{1}{k_B T}$ and $z(\vec{a}) = \sum_i \vec{a}_i \gamma_i$ gives the net charge of $\vec{a}$. By defining $\Delta E(\vec{a})$ as:

$$\Delta E(\vec{a}) = -2.3\kappa_B T \sum_i \vec{a}_i \gamma_i \left[\mathrm{p}K_{i,j}^{\mathrm{int}} - pH\right] + \sum_i \sum_{j \neq i} \vec{a}_i \vec{a}_j \Delta W_{ij} \tag{2.30}$$

we can express $p(\vec{a})$ as:

$$p(\vec{a}) = \frac{e^{-\beta\Delta E(\vec{a})}}{\sum e^{-\beta\Delta E(\vec{a'})}} \tag{2.31}$$

However, to properly estimate protonation states for a protein, we need to sample the several residues in a conformation using the Monte Carlo method and assign a protonation state to each residue. If the number of titrable sites in the protein is far too great, it becomes near impossible to calculate the $p(a)$ explicitly for every residue. Furthermore, not every configuration is eligible to be sampled. A certain criterion must be met:

$$P_{i \to j} = min\left\{1, e^{\Delta E_{\vec{a}1 \to \vec{a}2}(pH)/RT}\right\} \tag{2.32}$$

This Metropolis criteria establishes that a new protonation state is always accepted, if it has a lower protonation free energy than the previous protonation state. Otherwise, it will have a probability of $e^{(\Delta E_{\vec{a}_i \to \vec{a}_j}(pH)/RT)}$ to be accepted. These exchange attempts will happen, iteratively, for every residue and possible protonation state. After a number of MC steps, there will be a valid sample of possible protonation states to be evaluated and extract several properties, such as the pK_{a} value for the residue (if these calculations are submitted for different pH values).

2.4 The Constant-pH MD Method

With the tools for the job, the stochastic titration CpHMD method, developed by Baptista et al. [23, 25–28, 48, 49]], defines itself as the result of the cooperativity between two different approaches of sampling phase space (MD) and to sample protonation states (MC) from the semigrand canonical ensemble. Molecular dynamics allows time-dependent properties to be determined, but it is not ideal when it tries to calculate protonation free energies. On the opposite side of the spectrum, Monte Carlo methods are able to study events of protonation for rigid structures with higher convergence, even if it lacks the structural flexibility characteristic of MD systems. A united front of these two methods allows small MD simulations starting from MC generated protonation states to be sampled, thus combining the sampling while integrating the time-dependent properties.

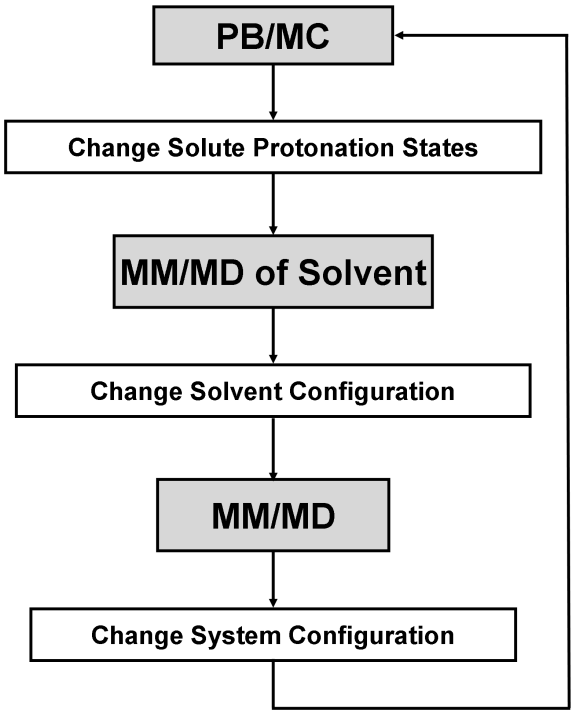

Figure 2.6: Simplified representation of the CpHMD method adapted from [27].

The presented scheme aims to describe, in a simple manner, the inner workings of the CpHMD method. The first step requires PB/MC calculations to estimate the protonation free energies, at a specific pH value, and to assign protonation states to the titrable residues. The chosen state is the selected one in the last MC step. Following the possible protonation changes of the solute, a small MM/MD step is performed with constraints placed upon the solute. These constraints will not allow movement of the solute, while the solvent molecules adapt to the new charge configuration, correcting the non-favorable interactions and establishing more electrostatically compatible positions around the solute. After this relaxation step, where the solvent and configuration is different from the initial structure, a longer MM/MD step will be performed. This effective step provides a full MD simulation of the system, during the length of the step, ensuring variability and conformational changes of the whole system while retaining the assigned protonation states for the residues. The last conformation obtained is the rigid structure for the following PB/MC step, initiating a new cycle.

2.5 pH Replica-Exchange

2.5.1 Sampling Enhancement: Replica-Exchange

The sampling problem has been previously delved upon in section 1.2.1, where temperature replica-exchange molecular dynamics (T-REMD) was briefly discussed. In this subsection, we aim to provide a slightly more detailed description of REMD in order to expand the concept to CpHMD. As previously mentioned, REMD is an enhanced sampling technique that allows higher energy states to exchange with lower states, facilitating jumps between minima while exploring the phase space. Each REMD run consists on multiple independent and simultaneous simulations called non-interacting replicas. Each replica differs from another based on a system parameter such as the temperature. Using temperature as an example, surpassing energy barriers are a result

of exchanging system states from higher temperatures with lower temperature states obtained from different replicas (i and j). Since it is possible to estimate the probability of a system state with an energy E based on the Boltzmann distribution, we can also estimate the probabilities of other system states at any given T temperatures.

A proper estimation of the probability of permutation between two replicas is crucial in a RE scheme. Allowing an exchange between two replicas, which is the same as exchanging two temperatures, it is possible to rescale the positions and momenta of both states according to a ratio between their new system temperature and the previous one. This rescale will affect the region of the ensemble which will be sampled and allow variability in both replicas. Though, the probability of exchange must respect the detailed balance condition for the transition probability (w). The detailed balance condition dictates that for a given state X for the system to transition to state X' must be equal to the probability of X' transition to X. In short, both transitions from state X to X' and X' to X must be accepted to properly exchange depending on their $w(X_{\rightarrow}X')$, which must pass the Metropolis criterion:

$$w\Big(X \rightarrow X'\Big) \equiv w\Big(X_m^{[i]}|X_n^{[j]}\Big) = \begin{cases} 1, & \text{for } \Delta \leq 0, \\ exp(-\Delta), & \text{for } \Delta > 0 \end{cases} \tag{2.33}$$

where,

$$\Delta \equiv \Big[\beta_n - \beta_m\Big]\Big(E(q^i) - E(q^j)\Big) \tag{2.34}$$

So in order for an exchange to be successful between two neighboring temperatures (n and m) from two different states (i and j) the probability of transition given by eq 2.31 must satisfy the Metropolis criterion.

2.5.2 RE applied to the CpHMD

The stochastic titration CpHMD method is successful in predicting pK_a values of titrable residues [23,25–28,48,49]. However, there are sampling issues related to the ability to predict this property in more challenging cases such as membrane systems. For example, residues that reside deep in a protein pocket or are close to membrane acyl chains, where water is scarce, will not provide sufficient sampling of charged states. Insufficient sampling of different protonable states in a group of conformations will lead to exaggerated Hill coefficients when fitting a titration curve and, ultimately, unreliable pK_a values or no values at all [37,38].

In order to increase sampling of both these conformational and protonation states, the pH replica exchange (pHRE) method is introduced. The pHRE theoretical foundation is sustained by the CpHMD method as presented in Figure 2.6. Due to the characteristics of the RE method, multiple independent CpHMD simulations are ran simultaneously while, in this work, allowing the pH of each simulation to be exchanged with another pH value. These exchanges are allowed within a certain probability which can be described as such:

$$P(X_i^l)P(X_j^n)w(X_i^l, X_j^m \rightarrow X_i^m, x_j^l) = P(X_i^m)P(X_j^l)w(X_i^m, X_j^l \rightarrow X_i^l, x_j^m) \tag{2.35}$$

where $P(X_i^m)$ is the probability of a certain protonation state m at pH value i, while X^{l} and X^{m} are protonation states from different replicas. The previous equation can be further derived until we obtain,

$$P_{i\to j} = min\Big\{1, exp\big[\ln 10(\mathrm{N_i} - \mathrm{N_j})(\mathrm{pH_i} - \mathrm{pH_j})\big]\Big\} \tag{2.36}$$

This last equation is a Metropolis criteria for the exchange probability where N_i is the number of titrable protons present in conformation i at pH_i. In our approach, the exchanges were attempted between neighboring values. The reasoning is that higher differences between pH values and/or number of titrable protons will exponentially decrease the probability of exchange. If, between two states, the number of protons is vastly different, the exponent in equation 2.34 is too large and the probability will be very low. Although this method relies only on attempting pH values exchanges, there are two other possible derivations from the initial detailed balance condition which may grant two other exchange methods: a protonation state exchange and a pH value **and** protonation state exchanges. If we attempted an exchange between both pH values and protonable states, we would be taking into account the potential energy for both states while losing the explicit influence of the pH values.

$$exp\big[\beta(\Delta V(q_i, N_j^m \to N_j^l) + \Delta V(q_j, N_i^l \to N_j^m)\big] \tag{2.37}$$

Since each protonation states comes from two independent replicas, their original protonation states are the most favored for the residues. Then, the energy required for a different protonation state to be accepted in a conformation where it would not be favored is insurmountable. On the other hand, if the exchange was only between the protonable states, the same problem would arise with the protonation exchange term contributing with an high energy barrier that the second term could not even out:

$$exp\big[\beta(\Delta V(q_i, N_j^m \to N_j^l) + \Delta V(q_j, N_i^l \to N_j^m) + ln10(N_i - N_j)(pH_i - pH_j)\big] \tag{2.38}$$

2.6 Simulations Settings and Parameters

This following section will present the settings and parameters used in the simulations that were performed with the previously mentioned methods. Across all the simulations, the studies will reflect on two pHLIP variants: *wt* and L16H, both in 1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) membrane bilayers. Before heading to more detailed information, a more general view will be presented. Regarding the CpHMD methodology, a total of 24 simulations of 100 ns each were performed. For each system, three replicates were simulated in the same parameters in order to statistically validate the results. In each replicate, the simulations were ran over 4 different pH values, ranging from 4.0 to 7.0 since these pH values confer different affinities for the residues regarding protonable states. Thus, allowing proper titration curves to study the insertion and residue $\mathrm{p}K_\mathrm{a}$ values. These studies will establish a cornerstone to all the following results and simulations setups.

For the pHRE simulations, all simulations were performed for 100 ns and they can be sub-divided in two groups: the 256 (group A) and the 128 (group B) POPC bilayer. Several simulations were performed with the purpose of determining which parameters are the optimal, regarding both speed and result validity. Another case of parameter comparison was related to the pHRE exchange frequency (τ_{RE}) attempts for which we tested 20 and 100 ps. As such, group A was composed of another two subgroups: the 20 and 100 ps τ_RE. Each subsystem comprised of three replicates, where each replicate had the same pH range as the CpHMD simulations (pH 4.0 to 7.0). Finally, group B ran simulations for both pHLIP variants (*wt* and L16H), each variant required 5

Table 2.1: Overview of all the combinations of methodologies, variants and parameters used in the simulations presented this thesis.

Methodologies	Variant	Number of Lipids	τ_{re} (ps)	pH range (step)	Number Replicates
CpHMD	*wt*	256	-	4.0 → 7.0 (1.0)	3
	L16H				3
pHRE	*wt*		100		3
			20		3
	wt	128	100	4.0 → 7.5 (0.5)	5
			20		5
	L16H		100		5
			20		5

replicates each, each replicate had 8 different independent simulations, whose pH values ranged from pH 4.0 to 7.5 with 0.5 pH value step. Again, different τ_{RE} values were used, using either 20 or 100 ps. This reflected on a total of 160 independent simulations for group B (*wt* and L16H simulations).

2.6.1 MM/MD settings

All the MM/MD simulations were performed with modified versions of the GROMOS 54A7 force field [17, 50] and the GROMACS 4.0.7 package [25, 26].

In all simulations of pHLIP, it was used the v-rescale thermostat [42] at 310 K with individual couplings for solute and solvent and a relaxation time of 0.1 ps. Regarding the pressure for the systems, it was used the Parrinelo-Rahman barostat [45] for a semi-isotropic coupling at 1 bar, with an isothermal compressibility of 4.5×10^{-5} bar^{-1} along with a relaxation time of 5.0 ps. For all simulations, the integrator timestep was 2 fs and in order to constrain all the solute bonds, a P-LINCS constraint algorithm [51] was used, while the SETTLE algorithm was used for water molecules (simple point charge, SPC [52]).

Long-range electrostatics were treated using a generalized reaction field [41] with a 0.1 M ionic strength and a relative dielectric constant of 54 [53]. Meanwhile, the non-bonded pair lists were updated using a twin-range cutoff scheme. With this scheme, the atom pairs which are inside the 0.8 nm cutoff were updated every step while the pairs between 0.8 and 1.4 nm are updated every 5 steps.

2.6.2 PB/MC settings

The Poisson-Boltzmann calculations were performed using the DelPhi software, version 5.1 [54, 55]. These calculations used the partial charges from the GROMOS54A7 force field and atomic radii derived from the Lennard-Jones parameters [56]. The solute molecular surface was defined using a probe of radius 0.14 nm [57] and an ion exclusion layer of 0.2 nm. The dielectric constants used were 2 and 80 for the solute and solvent [25, 26].

When performing the PB/MC calculations, a two step focusing procedure [58] is used. This procedure is based on the finite difference between grids points. The coarse grid has 91 grid points and, after a focusing step, the grid size is reduce to one fourth of the grid points in the coarser grid. This results in a spacing of 1.0 and 0.25Å between grid points in the coarse and focus grid, respectively. The maximum amount of linear iterations for these set of procedures was 100. This allows the maximum potential change to converge to it is threshold value (0.01 kBT/e). Finally, the relaxation parameters for the linear and non-linear forms of the PB equation were 0.2 and 0.75, respectively.

The energy values obtained from the calculations are computed, as previously said, in sampling protonation states using a MC procedure. The protonation states sampling was obtained using the PETIT program, version 1.6 [24, 59]. The MC runs consisted of 10^5 MC cycles, where each cycle attempts several sequential state trial changes over all individual sites and pairs of sites [24, 59] with an interaction larger than 2 units. The last protonation state is used in the following MD segment.

2.6.3 Simulation settings

All the CpHMD and pHRE simulations were performed with the stochastic titration constant-pH method [23, 25–28, 48, 49] (see section 1.3). Each CpHMD cycle was 20 ps long and each solvent relaxation step was 0.2 ps long, in both CpHMD and pHRE simulations.

For both CpHMD and pHRE, the protonation states of the titrable sites were allowed to titrate in the following pH ranges: from 4.0 to 7.0 (1.0 steps) for *wt*/L16H CpHMD and *wt* pHRE in 256 POPC systems; from 4.0 to 7.5 (0.5 steps) for *wt*/L16H pHRE in 128 POPC system. Initially, the systems were prepared with 256 POPC lipids, though they were further reduced to 128 lipids since it led to faster simulations and the results obtained were similar. The reason for the first 256 systems had the leeway to ensure that pHLIP would not see its periodic image, once it partially exited the membrane.

All simulated systems went through minimization procedures consisting on three sequential steps. The system's energy was minimized using the steepest descent approach without any constraints in a first instance, followed by 2000 steps of l-BFGS without constraints and a final run using again the steepest descent method with constraints on all bonds.

Following the minimization procedure, an initialization of the system was performed in two independent steps. The first independent step consisted on 100 ps of MD simulation with position restraints applied to all the atoms in the system. After this step, another 200 ps of MD is ran while applying restraints on all the C_α's atoms. A force constant of 1000 $kJmol^{-1}nm^{-2}$ was used in both steps.

2.7 Analysis

2.7.1 Secondary Structure - DSSP

The secondary structure analysis was performed using the DSSP program [60]. The DSSP program can be used in protein visualization tools, crystallography and NMR. This program can work in two ways: as a database of pre-existing PDB entries or as an application that will analyze a PDB file, which was the option used in this work.

By inputing a PDB file, the 3D structure of a protein can be deduced by reading the position of protein atoms in space while, at the same time, calculating the H-bond energy between all atoms. By placing the atoms, several specific angles and distances can be inhered. Thus, patterns will appear that will correspond to certain secondary structures, hence it will determine the most likely class of structure for every residue. These DSSP calculations were performed with the DSSP program which was called within the GROMACS 4.0.7 package.

2.7.2 Root Mean Square Deviation (RMSD)

Nowadays, the root mean square deviation (RMSD) is a simple and widely used method to evaluate and quantify the similarity between two structures by superimposing their atomic coordinates:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{2.39}$$

Where N is the total number of atoms, while δ_i is the distance between the atom i and the same atom in a reference structure. The RMSD program tries, in a first instance, to fit both structures by superimposing them through rotations and translations until it hits a minimum value. Furthermore, the RMSD calculations can be made for all atoms or a certain group of them. As such, we choose the C_α's atoms since the main structure will present less deviations since the main chain has less variation then the sidechain of the residues, thus lowering the RMSD value. The RMSD was calculated using the GROMACS 4.0.7 package.

2.7.3 Radius of Gyration

Similarly to the RMSD, the radius of gyration gives insight into the protein structure. The radius of gyration shows how compact a protein can be (eq. 2.38).

$$R_g = \left(\frac{\sum_i ||r_i||^2 m_i}{\sum_i m_i}\right)^{1/2} \tag{2.40}$$

Where r_i conveys the distance between a given atom i and the center of mass of the protein and m_i is the mass of that particle i. The radius of gyration allows us to see if there are deviations such as conformational changes and understand how the fold of the protein fits with the rest of the picture deduced for the protein structure.

2.7.4 Membrane Insertion Methods

The study of the pK_a of insertion (p$K_\mathrm{a}^\mathrm{ins}$) values of key residues such as Asp14 in pHLIP is one of the main objectives, thus we need to implement an accurate and robust measure of the insertion in the membrane. The idea is to obtain a measure of solvent exposure and a property that provides a gradient for the transition between water and the membrane. Ultimately, this insertion property will be a variable in the pK_a calculations and allow for measuring pK_a profiles along the membrane normal. With that purpose, three different methods were evaluated in the course of this work. The three methods share one crucial point: we measure the insertion of a given residue by following

one of its reference atoms along certain phosphour atoms of the membrane. The phosphour atoms used as reference will vary along with the method.

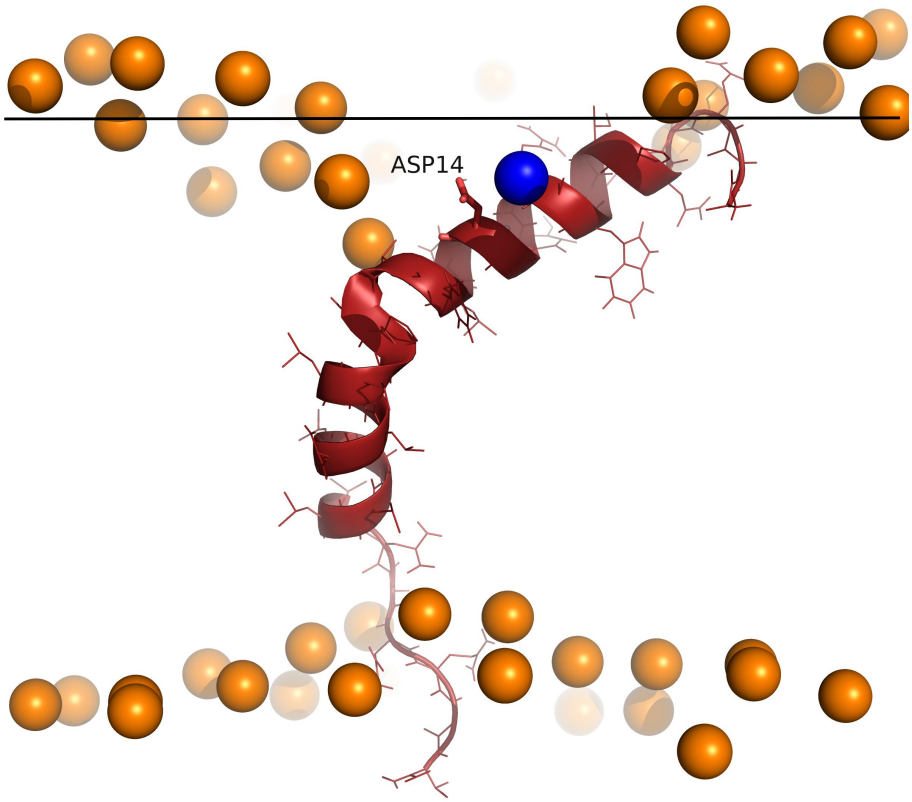

Figure 2.7: pHLIP peptide inserted across a POPC membrane bilayer obtained from a pHRE simulation. Phosphour atoms are depicted as orange spheres and the closest phosphorous atom to Asp14 is colored blue. The black line represents the average Z coordinate of the membrane.

The first method (the average phosphate method) used, as a monolayer reference, the average positions of the phosphate atoms of POPC. By taking into account each value for the Z coordinate in a *xyz* coordinate system, we can determine the average position of the phosphates for each monolayer. With this average, if we take the Z coordinate of the reference atom of the residue of interest, we can measure the relative distance between them and establish how inserted the residue will be. This is calculated for each frame for every trajectory.

The second method (the closest phosphate method) was based on the relative distance of the reference residue atom and the closest phosphate atom of the membrane. We consider a residue inserted if it is below the closest phosphate atom (fig. 2.7), with decreased solvent exposure, then we can measure the relative distances to every phosphate atom and determine which is the closest. After that, we can establish the extent of insertion of a residue relative to the selected phosphate.

Both methods have their merits and flaws. The average phosphate method, although it gives a good approximation of the relative Z position of the membrane as an average value, it fails to properly discern local deformations in the membrane and how they affect the relative position of the residue. For example, in figure 2.7, the average position of the phosphates appears to be too much above the Z position of the residue failing to provide an accurate measure of its solvation. Since there is a

depression in the membrane, the closest phosphate atoms are in line or slightly above the residue. Which seems to be a much beter description of the solvent exposure experienced by the residue of interest. Thus, the average phosphate method has a major flaw in dealing with such cases.

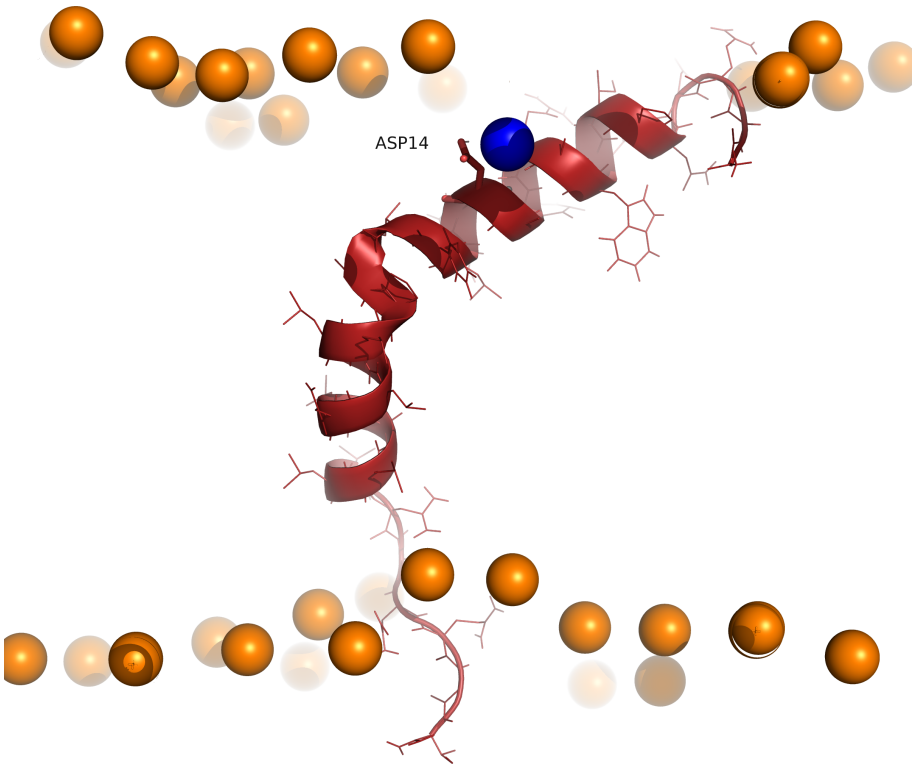

Figure 2.8: pHLIP peptide inserted across a POPC membrane bilayer obtained from a pHRE simulation. Phosphour atoms are depicted as orange spheres and the closest phosphorous atom to Asp14 is colored blue.

In figure 2.8, we present a scenario where the closest phosphate atom is in the same plane of the residue, hence the residue would be marked as significantly exposed to solvent. However, if this closest lipid is interacting strongly with the group of interest, they both can be dragged away or inside the membrane, rendering all these conformations to have a wrongly midpoint insertion. In these cases, one would get a better description with an average phosphate value. With no major deformation of the membrane, the 'rogue' phosphate error would be compensated by the average.

Both methods show flaws and advantages, which led us to propose the development of a third alternative. This new method uses as a reference, the average position of the phosphour atoms within a given cutoff. This cutoff needs to be large enough to gather several P atoms, but small enough to properly describe the membrane local deformations. A value of 6 Å gives usually very good results. In some snapshots, we might not find a P atom within the cutoff and, in that case, the closest phosphate atom will be used. The distance cutoff method is shown in the following figure 2.9.

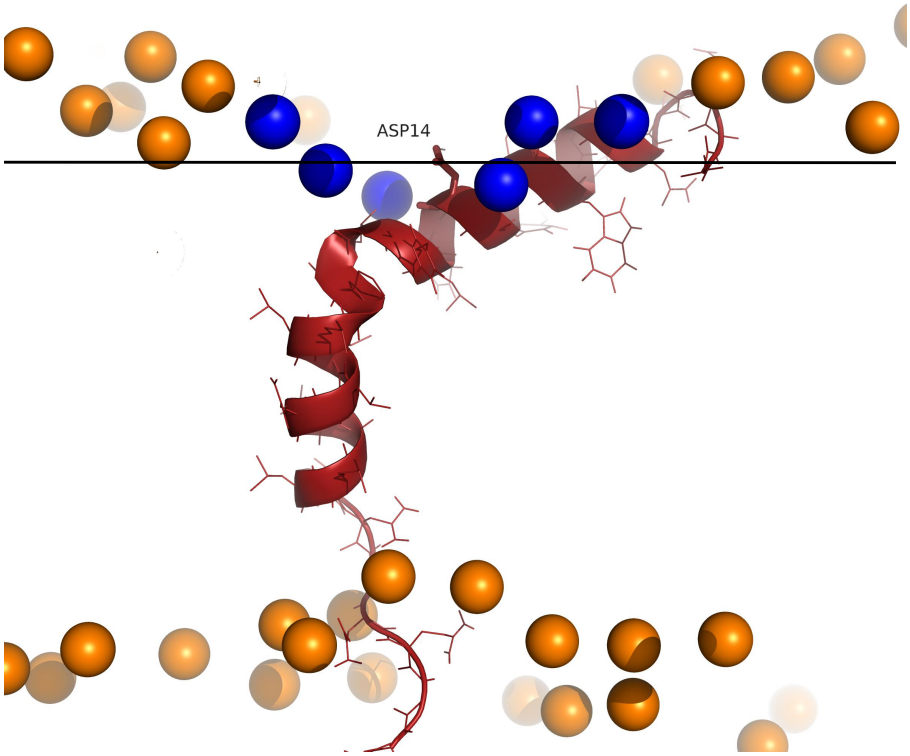

Figure 2.9: Graphical representation of the comparison between the distance cutoff and the average phosphate methods. The pHLIP peptide is inserted across a POPC membrane bilayer and represented in red cartoon. Phosphour atoms are depicted as orange spheres and the closest phosphorous atom to Asp14 is colored blue.

Using this method, it is possible to suppress the inherent flaw of the closest phosphate method by calculating an average of the Z positions and, since those atoms are within a short distance range of the residue, the depicted local environment will be more accurate for the residue. The method works by evaluating the relative distances of all the phosphate atoms to the reference atom of the residue. The phosphate atoms within a radius of 6 Å will be accounted for the average phosphate Z positions. Finally, the insertion value that is obtained for the residue is given by the relative difference of the Z coordinate of the reference atom to the calculated phosphate average.

2.7.5 Membrane Thickness

Along with the insertion studies and pK_a predictions, a thorough analysis of the membrane characteristics sheds light upon how accurate our simulations are being performed. The unperturbed membrane thickness can be evaluated using experimental results, thus validating how realistic these simulations and results are. Furthermore, due to the presence of charged residues, it aids in the evaluation of what kind of impact the peptide has on the membrane surface. Membrane deformation is directly correlated to the way the peptide rearranges itself when inserted. The spatial rearrangement not only depends on the surrounding environment but also on how the titrating residues interact with each other.

In this work, the membrane thickness was calculated analogously to the third insertion method (see subsection 2.7.4). When taking a snapshot of the system, we have the membrane bilayer with the pHLIP peptide inserted across it as previously represented (fig.1.1) . In that snapshot, the first step requires the center of the membrane to be calculated by identifying the lipids in the bilayer that are not perturbed by the peptide. These are usually the lipids beyond a XY distance cutoff of 15 Å. When identified, these lipids define the membrane center, which will be used as reference to estimate the individual monolayer thickness values. The bilayer thickness of the region beyond the cutoff (bulk) can also be compared with the experimental values, for validation.

The next step is performed with a sliding anulus of two cutoff values using *xy*-distances. These distances are measured between the peptide and the P atoms and use the average Z positions of the P atoms within the cutoff to generate thickness profiles when radially moving away from the peptide. In more detail, the hickness for each slice is calculated as the difference between the average of the Z coordinates of the phosphate atoms in that slice and the previously obtained Z average of membrane center. The thickness profiles obtained should give an idea of the impact of the peptide/protein on both monolayers, independently, and also a measure of what distance the deformation disappears and the thickness converges to its bulk value.

2.7.6 Estimation of $\mathrm{p}K_{\mathrm{a}}^{\mathrm{ins}}$ values

The method used to estimate residue insertion $\mathrm{p}K_{\mathrm{a}}^{\mathrm{ins}}$ values requires both insertion values and protonation states. The insertion data is important in order to provide a $\mathrm{p}K_{\mathrm{a}}$ profile along the membrane normal for each residue. This profile may provide information on: the $\mathrm{p}K_{\mathrm{a}}$ values for deeply inserted residues in the membrane, the effect of the dielectric on the $\mathrm{p}K_{\mathrm{a}}$ values along the membrane and possible correlation between the deeply inserted $\mathrm{p}K_{\mathrm{a}}$ values and the experimental $\mathrm{p}K_{\mathrm{ins}}$ values [9]. As previously mentioned for the CpHMD method, we can assign protonation states to each titrable residue for the full extent of the simulation. Assuming that the titrable residues are exposed to water, we can expect them to be exchanging between protonated and deprotonated states. Concerning each residue, the pH value of each simulation will influence the degree of each protonation state to a certain degree. Hence, for an anionic residue such as Asp or Glu ($\mathrm{p}K_{\mathrm{a}}$ of 3.65 and 4.25, respectively), protonated states are more probable at lower pH values. The presence of the membrane creates a new variable, the degree of bilayer insertion.

In these system simulations, for each residue we got the same number of data points for protonation states as frames in the simulation. Since it also exists a residue insertion value for each frame, it is possible to correlate insertion and protonation data into pairs. We then sort and slice the data pairs by insertion values, where every pair is grouped into an insertion bin. Every value, within a range of -1 to 1 Å from the insertion value, will be included in the same bin. Therefore, we can sample several protonation states for each bin, thus allowing an estimate of a $\mathrm{p}K_{\mathrm{a}}$ value per insertion bin. In order to obtain statistically relevant data points, we applied a criteria to our procedure that required a minimum of 50 events of each protonation state before performing a $\mathrm{p}K_{\mathrm{a}}$ estimate. After every data pair has been distributed to an insertion bin, we fit every data point to the Hill equation (eq. 2.39) while using an implementation of the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm:

$$f(\mathrm{pH}) = \frac{1}{1 + 10^{\mathrm{n}*(\mathrm{pH}-\mathrm{p}K_{\mathrm{a}})}} \tag{2.41}$$

Where n is the Hill coefficient and $\mathrm{p}K_{\mathrm{a}}$ is the residue $\mathrm{p}K_{\mathrm{a}}$ at that degree of insertion [61].

2.7.7 Error Analysis

All pK_a error values presented in the profile plots were obtained by using a modified jackknife resampling method while, for the other results, it was estimated a standard error of the mean between all the replicates. A statistical resampling method proves to be a feasible way for model validation and hypothesis testing. A resample method works by drawing repeated samples from the original data in a randomized way allowing different combinations of cases, thus creating new sampling distributions from the original data. With new and different sampling distributions, we can yield unbiased estimates as long as the original samples themselves are unbiased. [62]

The used jackknife method requires combinations of the replicates that constitute the sampling for a given system. These combinations are achieved by deleting one replicate from the original sample, generating a different sample. Thus, if we have n replicates of a given simulation, to estimate the standard error we require all the sample combinations constituted of n-1 replicates. For example, if we have 5 replicates, the combinations of replicates that provide the samples are: $i = \{1234, 1235, 1345, 1245, 2345\}$. Analogously to the method used in [63], each combination was fitted to a Hill equation (eq. 2.39).In the following equation, we can observe the generalized formula of the used jackknife method to calculate the standard error(SE):

$$SE = \left\{ \frac{n-1}{n} \sum \left(\frac{\bar{x} - x_i}{n-1} \right)^2 \right\}^{1/2} \tag{2.42}$$

In the equation, n is the number of simulation replicates, $\bar{x}$ is the predicted pK_a value using all the replicates sampling and x_i is the pK_a value using the sampling in each i combination. The error bars plotted with the pK_a values are obtained with the standard error and they need to fulfill two criteria: each i combination must have at least 50 points for each protonation state and the sample points must come from, at least, 2 pH values.

Chapter 3

Results and Discussion

3.1 CpHMD Simulations: *wt*-pHLIP

With constant pH molecular dynamics (CpHMD) simulations, it is possible to study the molecular details of the pH-dependent membrane insertion of the pHLIP peptides. A first general overview of the structural features of pHLIP is made in order to further develop into the intricacies of the insertion process.

3.1.1 Conformational Analysis

The secondary structure using the DSSP criteria

The *wt*-pHLIP states, in the mechanism of action, are well defined from an experimental perspective. Even though the insertion process is not well understood, we know that the transitions between state II and III are coupled with the insertion and withdrawal processes within the milisecond to the second timescale. As such, it is not possible, with our methodology, to observe these transitions, because we can only sample those within the nanosecond timescale. Nevertheless, since state III stability is a key determinant in the whole process, we can study the pHLIP peptides in the membrane at varying pH values. The transition from state III to II is triggered by an increase of the pH value, which leads to ionization of the anionic residues and can have an effect on the structure of the peptide or deformations in the membrane.

Evaluating the structural impact of different pH values is a challenge. The way that pHLIP accommodates in the membrane is heavily dependent on electrostatic interactions between the key residues, interresidue interactions with the membrane and solvent exposure. All these subtleties add to the variety of local energy minima that can be explored, which further increases both the variability and the complexity in sampling this peptide.

The secondary structure of *wt*-pHLIP was obtained for every simulation performed with CpHMD in a 256 POPC lipid bilayer. In Figure 3.1, it is depicted the secondary structure that each residue adopted throughout the simulation time at the highest and lowest sampled pH values.

In a first instance, we can observe that there is an α-helix region comprised of the residues 5-33 in both representations. Additionally, the residues 15-20 show a small turn which indicates the presence of an helix distortion in that region probably due to the presence of a proline residue. Partial exiting with loss of helicity may be associated with a metastable transition state of pHLIP

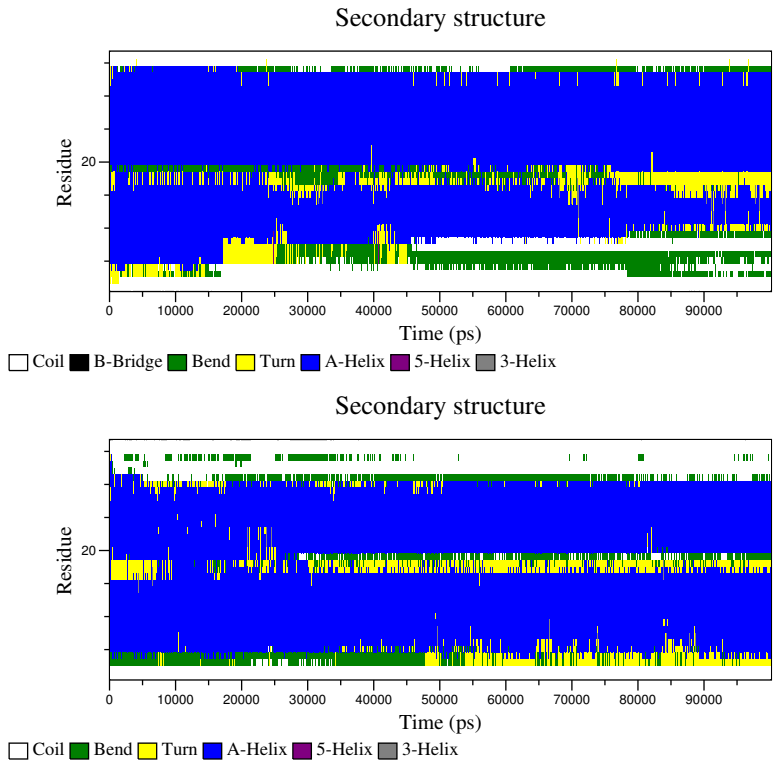

Figure 3.1: Representation of the secondary structures observed in state III pHLIP. These representations were obtained from simulations of *wt*-pHLIP in a system with 256 POPC lipids performed with CpHMD at pH 4.0 (**top**) and pH 7.0 (**bottom**) .

since, in state II, it does not possess such well defined secondary structure (coiled state) while, in the third state, pHLIP adopts an α-helix. If these meta-stable conformations of pHLIP are correlated with the withdrawal process due to the higher pH values inducing a loss of helicity, the next logical thought would be: what is the effect of pH on the secondary structure.

In Figure 3.2, there is no semblance of correlation between the pH value and the percentage of helicity of the residues. The average values of the three replicates show that the percentage of helicity at different pH values are within the standard error. Therefore, we can not state that these meta-stable conformations are dependent on the pH nor if they are associated with the insertion/withdrawal process.

The linear response approximation (LRA) was the first used method to predict the pK_a values of *wt*-pHLIP MD simulations. Each MD simulation had fixed protonation states, hence it was required to sample the two possible protonation states of Asp14 to predict pK_a values. Thus, the system was simulated in two groups with all titrable residues in the charged state with the exception of Asp14: protonated (group 1) and deprotonated (group 2). From the LRA simulations, we can observe that the helical content of *wt*-pHLIP is close to the CpHMD results. In the LRA methodology, we force the system to sample high energy protonation states which are, on average, less favorable. This premise of the method may lead to conformations where pHLIP does not fold into an α-helix due to electrostatic interactions between residues in unfavorable states for an α-helix. Therefore, we can conclude that the deprotonated states will present a lower helicity than *wt*-pHLIP with protonated states. In fact, the DSSP results of the protonated states are within the estimated error of the CpHMD results.

In conclusion, the aforementioned CpHMD DSSP results are consistent with the LRA results and

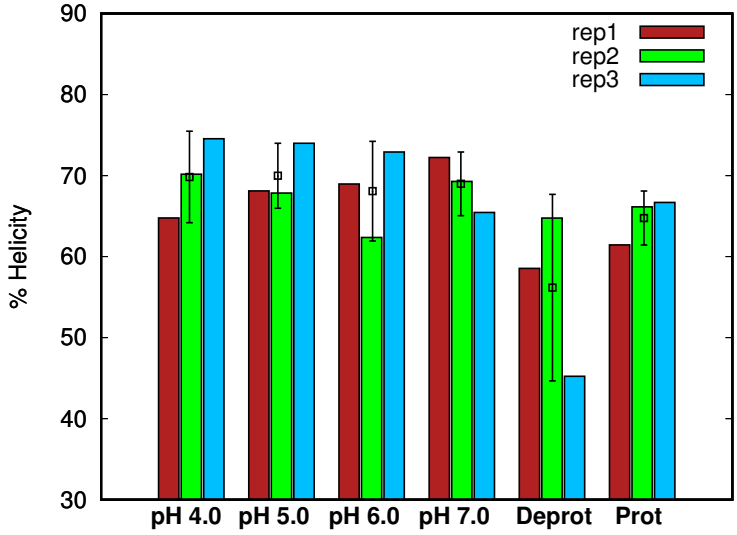

Figure 3.2: Percentages of helical content for *wt*-pHLIP LRA and CpHMD simulations. The pH values refer to the CpHMD simulations, while **prot** and **deprot** correspond to the protonated and deprotonated states of Asp14 in each LRA simulation [40]. The simulations were performed in a 256 POPC lipid bilayer. The black point in the graph represents the average percentage of the three replicates.

between themselves throughout the range of pH values and replicates. Hence, we observe that pHLIP is stable during the simulation, retaining its expected structure.

Thickness

The membrane thickness is an important property, which gives us information on lipidic phases and on local membrane deformations. In this work, we establish the assumption that a consistent value of thickness when moving away from the peptide, along the *xy* plane of the monolayer, indicates an unperturbed lipid bilayer. In the present study, we have a peptide of considerable size inserted across the membrane, so we expect some local deformation (Figure 3.3). These results show the thickness profie, in each monolayer, for the lipids that are radially distributed away from the peptide. At longer distances, we observe that the values converge to a plateau where the impact of pHLIP in the membrane is reduced. Henceforth, the region beyond the 20 Å will be considered a bulk region which could be compared to experimental results. The thickness of the bulk region for each monolayers converges to an approximate value of 20 Å and an overall membrane thickness of 40 Å. This value is consistent, within the standard error (Figure 3.3 lower panel), with the experimental value of 39 Å [64].

Figure 3.3 **top** shows the results of the N-terminus region, where Asp14 is inserting, and we observe that the peptide only has a small perturbation on the upper monolayer. There is a slight deformation of the membrane at 0-5 Å distances though it is not pH dependent. Membrane deformation is important since it defines the space where water molecules are allowed to reach and, consequently, which residues are exposed to solvent.

The membrane thickness in the peptide region must be evaluated for each monolayer separately, due to the distinct environment that affects each one. Figure 3.3 (**bottom**) shows the results at the C-terminus region. There are four anionic residues very close to each other (Asp31, Asp33, Glu34

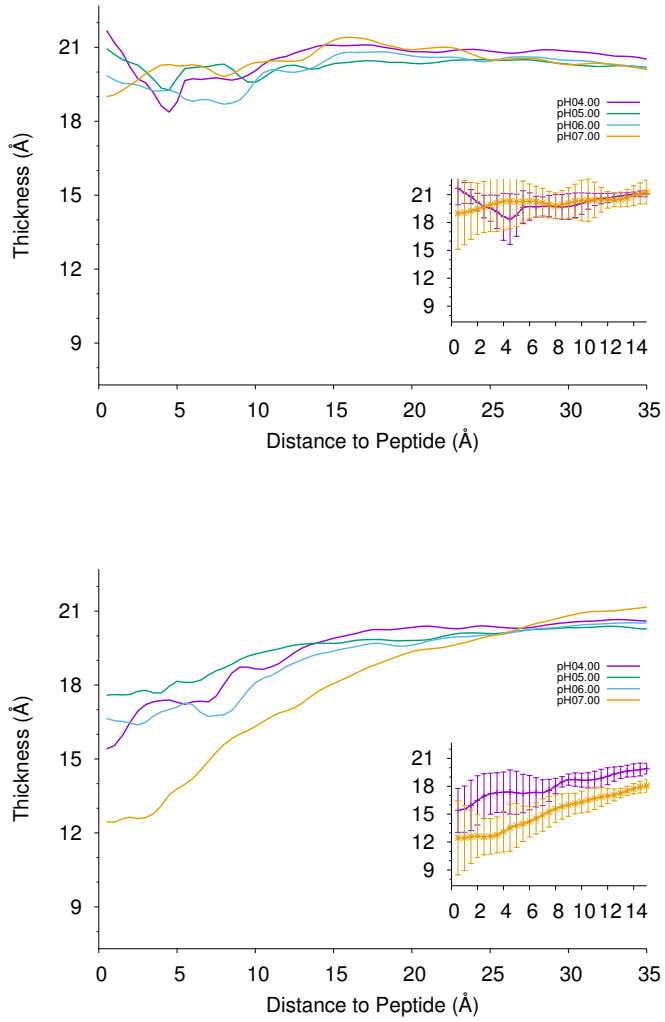

Figure 3.3: Representation of the membrane thickness profiles for the upper monolayer, the one interacting with Asp14 (**top**) and the bottom monolayer, the one interacting with all residues near the C-terminus (**bottom**). These representations were obtained from simulations of *wt*-pHLIP in a system of 256 POPC lipids performed with CpHMD. In the insets, we show the thickness with error bars for the higher and lower pH values studied. The error bars of the other two pH values are of similar magnitude and were omitted for clarity.

and Cter) at the C-terminus region. These anionic residues, when charged (higher pH values), induce larger deformations in the membrane. At pH 4.0, we expect these residues to be protonated, thus the electrostatic repulsion with the phosphate headgroups and the need for solvation is diminished. However, at pH 7.0, the charged states are more probable, hence a more negative environment ensues in that region, generating the electrostatic repulsion previously mentioned and significant local deformation in the monolayer, with water completely solvating the charged residues.

3.1.2 Insertion effect on Protonation

Membrane deformation will affect the insertion of key residues and will impact the pK_a estimations. In Figure 3.4, there are pK_a profiles varying along the membrane normal for several key residues. The results obtained for the residues in the C-terminus are relative to the transitions from state III to state II. Knowing that these residues pK_a values depend on the environment upon insertion, which is different in the two states we can only evaluate the peptide transition from state III to state II. On the other hand, Asp14 enters and exits the same monolayer in both transitions (II $\rightarrow$ III and III $\rightarrow$ II), which we can capture in our simulations.

We assume that the pK_a values of at least one the acidic residue is defining the experimentally observed pK_{ins}. The most inserted pK_a values calculated for these residues should be the best estimations for this pK_{ins}. The values measured at those insertions, despite being at the limit of our sampling, correspond to regions where the residues are poorly exposed to water, being at the limit of exchanging protons and feeling the solution pH value. Experiments have shown that one (or both) of the two Asp residues located in the transmembrane region of the peptide (Asp14 and Asp25) is responsible for the trigger protonation leading to membrane insertion. However, our results (Figure 3.4) show unequivocally that Asp25 preferred location is completely buried in the membrane, always protonated. Therefore, Asp14 is indeed the key residue in the pHLIP insertion process and its pK_a value at a deep region of the bilayer, should define the observed experimental pK_{ins}. Figure 3.4 **top** shows a clear correlation between pK_a values, obtained with CpHMD and LRA, and the experimental pK_{ins} [40]. The experimental pK_{ins} of *wt*-pHLIP is 5.96 [9] while, Asp14 at the deepest measured insertion, has pK_a value of 5.75 calculated with LRA and 5.69 from CpHMD. Both pK_a values are in good agreement with the experimental pK_{ins}, further validating the computational methods in predicting the pK_a values for pHLIP.

We should expect the pK_a variation along the insertion to be consistent with the typical profile of a carboxylic acid [65]. The aforementioned results were obtained using the previously described slicing method (see section 2.7.4). Due to the low membrane dielectric and the scarcity of solvent molecules, the ionized form of the carboxilic acid will be destabilized, so its pK_a value will get higher as the residue gets deeper into the membrane. Overall, we can observe this trend and the pK_a values of the residues increase with insertion in the CpHMD simulations. The LRA pK_a profile curve for Asp14 exhibited an unexpected behaviour, which is due to the biased sampling induced by the simulations at constant protonation [40].

The slicing method generates valuable information, but also creates complexity in the data analysis. In Figure 3.4 (**bottom**), we have 4 different pK_a profiles which were calculated on the exact same conformational ensemble. However, the conformations used in the slices of same depth for these residues will be different because every snapshot can have each residue at a different level of insertion. A consequence is that, even though these residues protonation states are highly correlated, the pK_a profiles are almost independent and should only be compared carefully.

The higher pK_a values of Asp33 and Glu34 compared with C-ter is a direct consequence of the

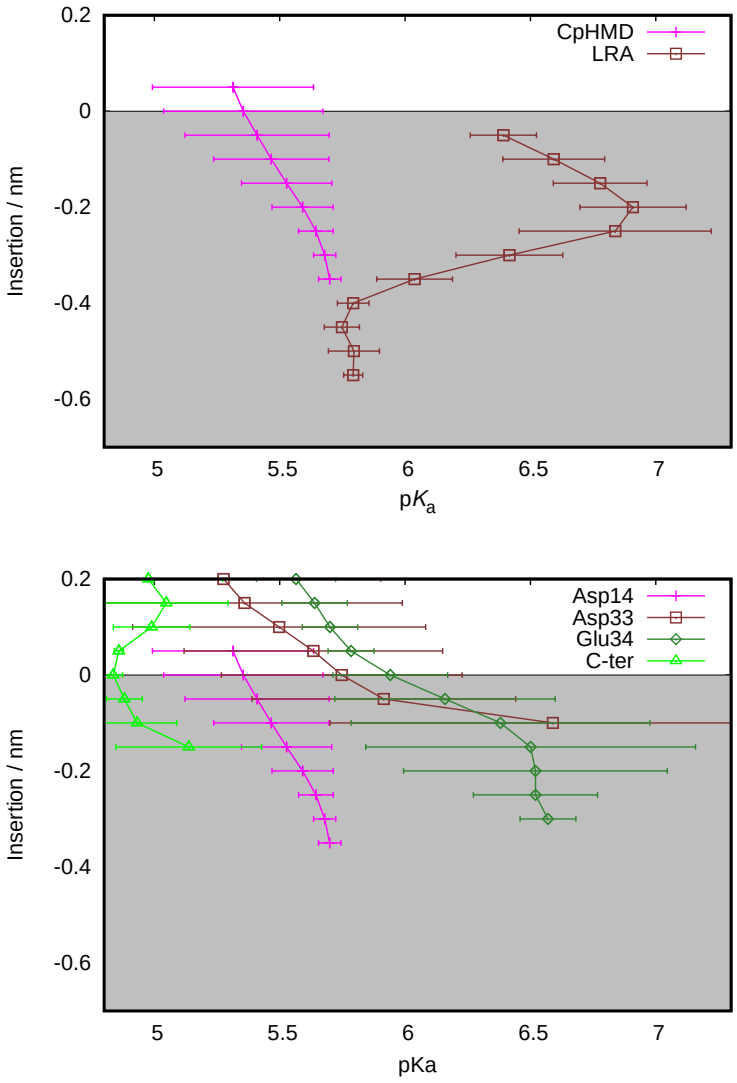

Figure 3.4: Asp14 pK_a profiles obtained with CpHMD and LRA(**top**) CpHMD key residues profiles (**bottom**). Key residues (Asp14, Asp33, Glu34 and C-terminus) pK_a profiles obtained from CpHMD simulations. The pK_a profiles are calculated along the membrane normal. Negative insertion values represent deeper membrane regions, while positive values represent water solvated regions.

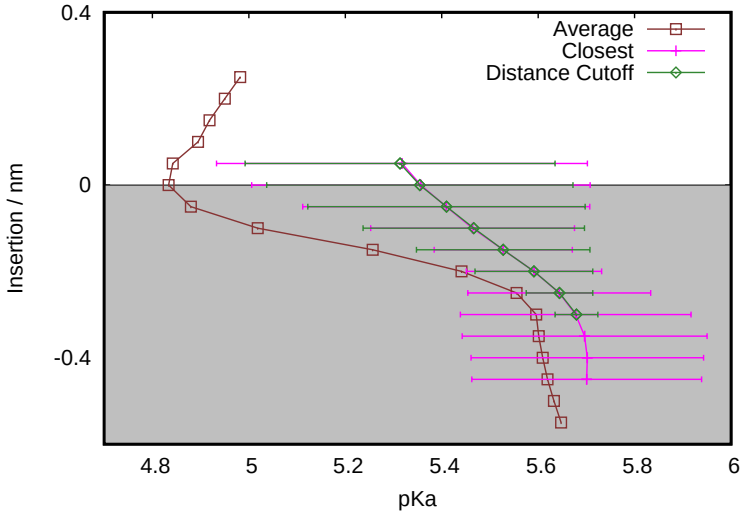

Figure 3.5: pK_a profile of Asp14 obtained using different insertion methods: average, closest and distance cutoff methods. The simulations were performed using the CpHMD method.

insertion process. Note that the conformations selected for Asp33 in a deep location will inevitably have Glu34 and C-ter in a more shallower region, which means that they will tend to be more ionized. The negative electrostatic potential created by these neighbors will push the pK_a values of Asp33 to even higher numbers. When Glu34 reaches the same region, Asp33 will be already deeper and neutral, so the electrostatic potential affecting Glu34 will be smaller, and the pK_a shift will not be as large as the previous. This also suggests that the withdrawal process is regulated by the cooperative protonation of all anionic residues in the C-terminus. By analyzing the error bars of Asp33 and lack of data of Asp31, we can state that the CpHMD method is able to predict the pK_a values of the residues to a certain extent but it lacks sampling of ionized states for the residues in the deeper regions of the membrane.

Significant concerns exist regarding the determination of the insertion values since, as seen in section 2.7.4, there are situations where the insertion is not properly evaluated. Previous results have shown that, depending on the used insertion method, we may observe significant disparity between the calculated pK_a profile for the same residue. Since there is a small, yet evident, deformation in the upper monolayer (Figure 3.3), we can infer that Asp14 may be inserted in that region and with its protonation state dependent on solvent exposure. Figure 3.5 shows the pK_a profile for Asp14 calculated with the three different insertion methods.

For the *average phosphate method*, the estimated pK_a value (5.64) is shifted towards the expected high pK_a values though unable to provide a reliable error estimation. Taking into consideration that Asp14 may be placed in the deformation, any insertion value will be falsely estimated because it only considers the average phosphate positions. With this method, Asp14 will appear to be in more inserted regions of the membrane, since the average Z coordinate, used as reference, will be higher than the local deformation. Therefore, well solvated ionized Asp14 conformations will be contributing to the pK_a calculation in deeper membrane regions where solvent molecules should be scarce, affecting the pK_a estimation. The pK_a profile using the *closest phosphate method* shows large error bars at each data point. When we use the *closest phosphate method*, we are overestimating the relevance of the position of a single phosphate atom. If a phosphate atom interacts with a residue and the residue begins to insert more, the phosphate atom will be dragged and the method will consider that the insertion value will be close to 0, even if it is not in contact with bulk water. The large error bars suggest that there is a mix of conformations that are considered

to have the same residue insertion, when, in fact, the residue is not in the same environment. This divergence in environment will lead to a dispersion in pK_a values for the same insertion value and we obtain the aforementioned pK_a profile.

We observe significant improvements on the profile and the error bars when using the new developed method. The new pK_a profile are similar to the pK_a profile of *the closest phosphate method*, but with a significant reduction in the error bars due to a better evaluation of the membrane surface. The new method represents the best of the two approaches, it mimicks the *closest phosphate method* in the attempt to capture the local deformations in the membrane, but still tries to use an average position from several phosphates atoms (within a given cutoff), in order to avoid solvent exposure heterogeneity in a given slice, due to strong interactions between the titrable site and the nearest phosphate. Henceforth, every depicted pK_a profile was obtained using the insertion values provided by the new slicing method.

3.1.3 *wt*-pHLIP Sampling Limitations

The LRA and CpHMD methodologies are able to estimate the pK_a value of *wt*-pHLIP Asp14 with a good agreement to the experimental pK_{ins}. However, we have seen that those merits fall short when we assess both the LRA pK_a profile (Figure 3.4) and the lack of sampled ionized states with the CpHMD. As we mentioned, the LRA pK_a profile does not match the typical pK_a profiles of anionic residues [61] and the CpHMD methodologies present consistent pK_a profiles but only with few data points, leading to large error bars, for a small section of the membrane. Some residues show a lack of ionized states to fullfill the imposed criteria, hence an absence of detailed profiles for Asp31 and Asp33.

In pHLIP simulations, the CpHMD methodology does not allow to jump energy barriers and sample high energy states. These high energy states may be unfavorable protonation states such as ionized states deep in the membrane. This restriction only allows sampling of more favorable protonation states, thus there will be less events of ionized residues in the membrane, culminating in poor pK_a profile estimation. Nevertheless, we expect that with better sampling of high energy states, we can describe deeper membrane regions and obtain more accurate pK_a values. Here, we will propose the use of an enhanced sampling method based on the CpHMD methodology - the pH replica exchange (pHRE).

3.2 pHRE Simulations: *wt*-pHLIP

3.2.1 Optimizing System Setup and pHRE Parameters

Using the pH replica exchange (pHRE) method, we should be able to sample protonation states in deeper membrane regions, thus improving the pK_a profiles for the *wt*-pHLIP residues. The pHRE requires, as mentioned in section 2.5.2, exchanges between different pH values since the conformations will possess certain protonation states that are not as favored (higher energy) as in their "original"pH value. The τ_{RE} is an important parameter in pHRE because it defines the rate of pH exchange attempts between replicas. We simulated systems with pHRE using two different τ_{RE}: 20 ps and 100 ps.

We started the CpHMD simulations using a pre-equilibrated 256 POPC lipid bilayer so there would not be any issue with pHLIP interacting with its periodic image, in case the simulation box was too small. As we observed in the thickness results for the CpHMD simulations (Figure 3.3),

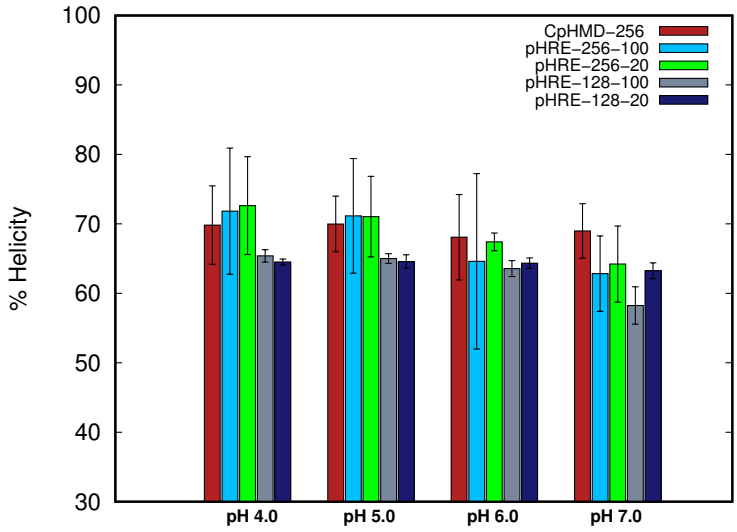

Figure 3.6: Percentages of helical content for *wt*-pHLIP using CpHMD and pHRE with different parameters.The simulations were performed with 256 and 128 POPC lipids bilayer. The τ_{RE} value used was either 20 or 100 ps.

we have a good margin of bulk membrane that is not affected by the peptide. Since the pHRE requires multiple independent simulations of several pH values, it is in our interest to increase the simulation speed. Therefore, we studied the impact of using 128 POPC lipids opposed to 256. These two parameters (τ_{RE} and number of lipids) are essential to optimize the simulations, before extending our study to the L16H pHLIP variant.

After performing simulations of *wt*-pHLIP using the two methods and varying parameters, we applied the same CpHMD tests to the pHRE simulations. At first, we studied the percentage of helicity in the four different pHRE groups. The results are comparable between the different pHRE simulations and in agreement with CpHMD results for all pH values. Therefore, we confirm that the peptide, in those simulations, remains, on average, at state III and we can safely estimate pK_a values without the interference of unstructured conformations.

Figure 3.3 shows, we have a considerable region of bulk membrane that stabilizes at $\sim$20 Å per monolayer, hence, there is a safe margin to reduce the size of the POPC membrane bilayer in the simulations. This downsize confers an advantage in simulation speed when compared with simulations with 256 lipids. Though that raises a question: how much can we reduce without affecting the peptide membrane interaction. The peptide accommodates in the membrane and they are in a equilibrium. If we reduce it too much, not only we fail to reproduce a viable pHLIP system but we can not measure reliable insertion values.

Figure 3.7 shows the thickness of each monolayer converging to approximately 20 Å for the 256 lipids pHRE simulations. The sum of both monolayer values is in agreement with the CpHMD and corroborated by the experimental value (39 Å) of a pure POPC lipid bilayer [64]. Regarding the pHRE simulations with 128 lipids, the monolayer thickness values do not converge near the end of the box. The downsizing of the membrane resulted in few bulk lipids and possible periodic effects. Nevertheless, the small estimated error values show consistency between the different pH values and the profiles suggests that it would eventually converge to the experimental value as seen in the simulations with 256 lipids. Furthermore, the local deformations in the membrane

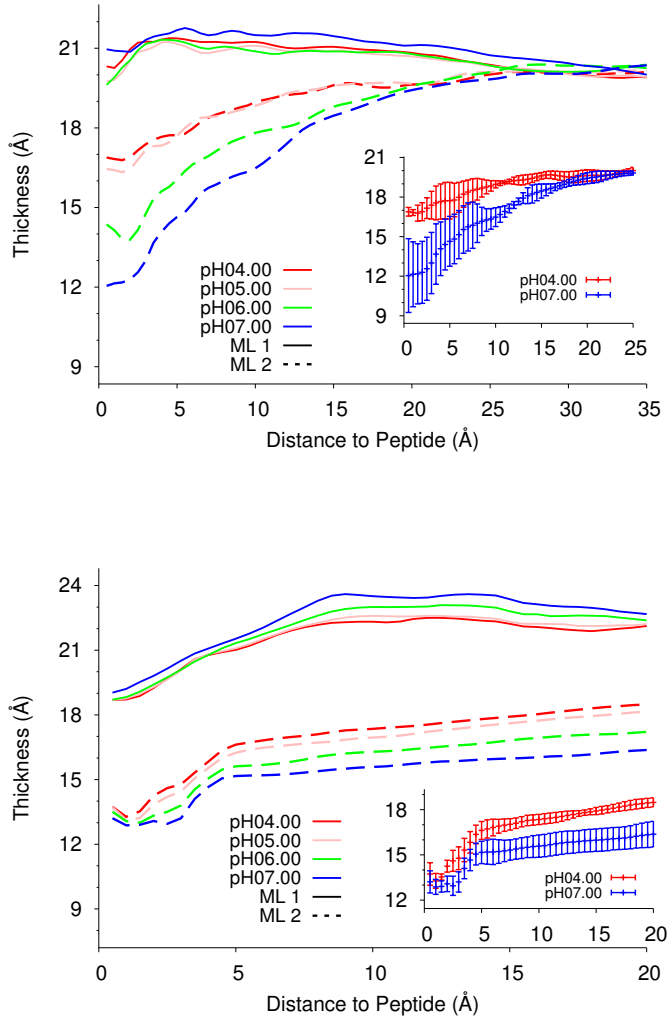

Figure 3.7: Thickness profiles for both monolayers (ML1 and ML2) in 256 lipids (**top**) and 128 lipids (**bottom**) pHRE simulations. Each pH value is represented without error bars for clarity. Some error bars estimations are shown in the smaller inset.

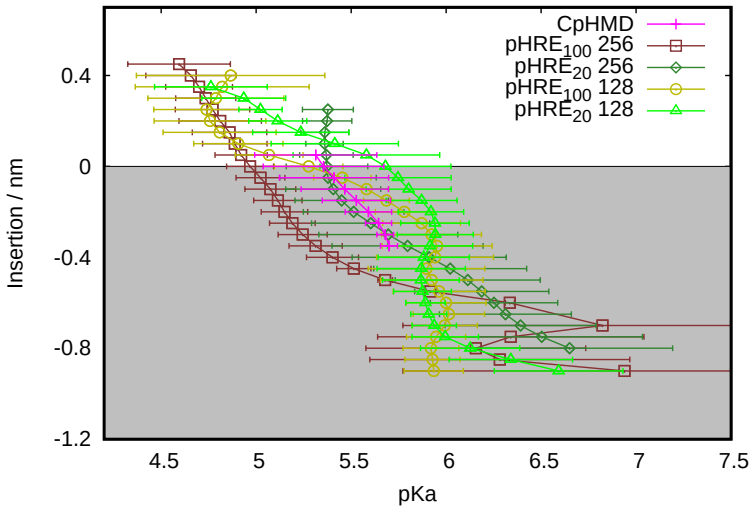

Figure 3.8: pK_a profiles for Asp14 in *wt*-pHLIP obtained from CpHMD and pHRE simulations with different τ_{RE} (20 or 100 ps) and number of lipids (128 or 256).

are similarly described, and this should be the most important factor influencing the pK_a profile estimations. Our rationale to reduce the membrane size remains valid, however a midway value, 160-190 lipids, would be more appropriate.

The quality of our results should also be evaluated when we change two other simulation variables: the τ_{RE} value (20 ps and 100 ps) and the number of replicas (4 or 8). In order to establish a reference τ_{RE} value and number of pH values to simulate, we compared pK_a profiles of Asp14 and probabilities of exchange between pH values for several pHRE simulations (Figure 3.8 and 3.9, respectively), we observe that, for deeper insertions in the membrane there are more ionized states sampled, thus fulfilling the required criteria and leading to more points in this region. Despite the different set of parameters, all pHRE simulations provided a significant improvement in sampling as we were able to obtain pK_a values for more deep and also shallow regions of the lipid bilayer. This happens in pHRE because a replica, at a given pH value, sampling a preferred region, will exchange pH value and provide conformations and protonation states for this new pH value that would be very improbable to sample. These extra points are pivotal to obtain more complete pK_a profiles in pHRE.

We observe that the pK_a profiles follow the expected carboxylic acid trend [61]. Even though the profiles *per se* differ, there is an observable solvent effect shared between every simulation. This solvent effect shifts the pK_a to higher values as the water molecules become more scarce and corroborating the large energy required to ionize a residue in that environment as it was observed in the CpHMD results. These results suggest that the pHRE simulations do not suffer from a memory effect. A memory effect can occur when two replicas exchange pH values and τ_{RE} is too short hindering a proper adaptation to the new pH value. This occurrence may create a bias in the estimated pK_a values and, consequently, in the residue profile. The bias effect is not observed in our simulations since the results are consistent between τ_{RE} values. These results are corroborated by the work of my colleague Pedro Reis regarding pentapeptides and tetrapeptides simulations with pHRE where he did not observe any memory effect in simulations with τ_{RE} of 20 ps.

The pK_a values at the deepest insertion value (Table 3.1) do not provide sufficient information to infer directly the quality of the method since the estimated pK_a values are almost all, within the error, in agreement with the experimental pK_{ins}. We can only deduce that both methods are equally

Table 3.1: pK_a values of *wt*-pHLIP Asp14 obtained with different approaches at the deepest measured insertion value. These should be a good estimation of the experimental pK_{ins}

Simulation	**Number of Lipids**	τ_{RE}	pK_a Asp14
Experimental	-	-	5.96 $*$
CpHMD	256	-	5.69 ± 0.04
pHRE		100	6.93 ± 1.17
		20	6.65 ± 0.54
	128	100	5.94 ± 0.16
		20	6.59 ± 0.34

$*$ This corresponds to the experimental pK_{ins}

valid in predicting the pK_a values of pHLIP. And the pK_a profiles show that the pHRE results are an improvement over the CpHMD method (Figure 3.8).

The probabilities of exchange between the two τ_{RE} do not depend on the frequency of the attempt (Figure 3.9). This is an important observation indicating that pH exchanges are being properly accounted for. With the same probability of exchange, simulations attempting an exchange every 20 ps will have more frequent exchanges than the 100 ps group. This suggests that attempting an exchange every 20 ps may prove to be more beneficial to shuffle pH values between replicas, further enhancing our sampling. Additionally, it is more beneficial to run simultaneous simulations with a ΔpH of 0.5 instead of 1, in these systems. Considering the Metropolis criteria of the pHRE (equation 2.36), the second term of the exponential depends on the jump between the pH values, thus higher pH jumps will decrease the probability of exchange. When using a ΔpH of 0.5, there is a considerable increase in probability of exchange. Hence, it is more advantageous to perform simulations with 8 pH values with a 0.5 step since, in the end, it increases the amount of exchanges, despite being more computationally expensive.

Ultimately, our results suggest that, for the pHLIP system, we can simulate pHLIP in a lipid bilayer of 128 POPC with a relative speed gain per independent simulation. Attempting pH exchanges every 20 ps further increases the accepted events as well as using pH values with a 0.5 pH value step. These two parameters may allow us to leave the local energy minimum and further explore the phase space.

3.2.2 pK_a Profiles of C-terminus Residues

We also performed pK_a estimations for the C-terminus residues (Asp31, Asp33, Glu34 and C-ter). Although, in our setup, these residues are not in the insertion process, but rather in a protonation-dependent withdrawal mechanism..

The pK_a profiles for the referred residues (Figure 3.10) show, at first sight, that pHRE seems to have solved a relevant problem of the CpHMD method which was the lack of data regarding the Asp31 and Asp33 residues. There was a deficit on the number of required ionized states to estimate a pK_a value due to the relatively deep zones they reside in. At those regions, the residues only ionize at pH values closer to 7.0 due to the desolvation effect and the low dielectric membrane shifting the pK_a values. Simultaneously, we are presented with an unexpected result. Although with few data points, the pHRE method was able to sample a sufficient amount of ionized states

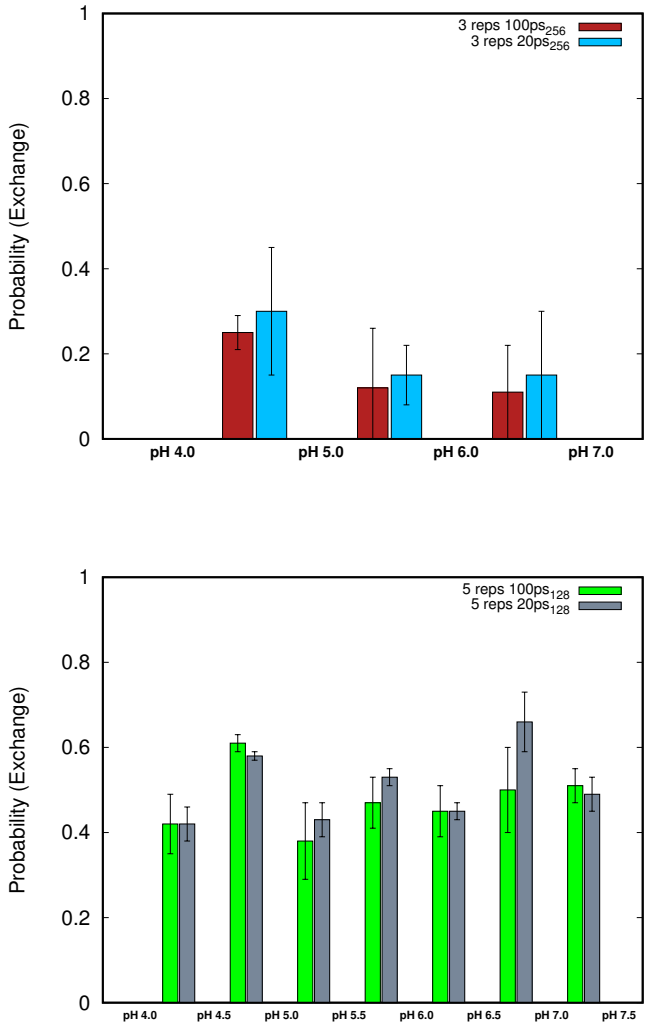

Figure 3.9: Probabilities of exchange between different sets of pH values. The systems with 256 lipids (**top**) were simulated from pH 4.0 to 7.0 (1 pH steps). The systems with 128 lipids (**bottom**) were simulated from pH 4.0 to 7.5 (0.5 pH steps). The plot bars are centered between the two interchanging pH values.

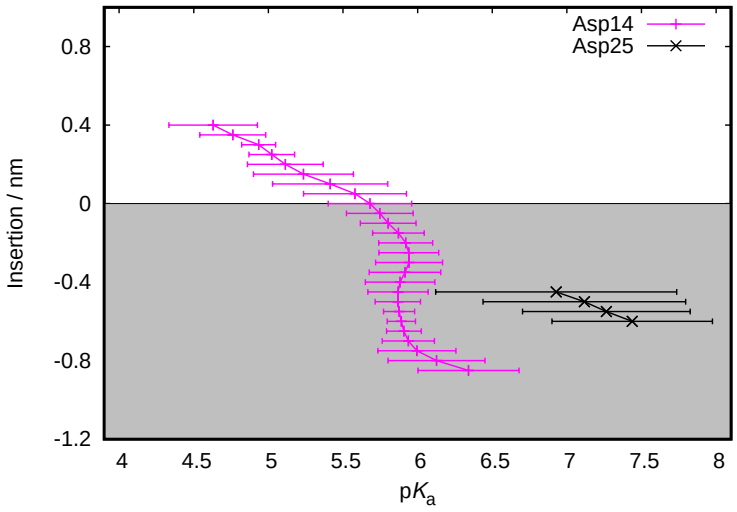

Figure 3.10: pK_a profiles for key residues of *wt*-pHLIP obtained in pHRE simulations with 20 ps τ_{RE} and 128 POPC in the membrane.

to fulfill our criteria, thus predicting some pK_a values for Asp25. The number of water molecules penetrating the membrane reaching Asp25 is small, therefore, we were never able to predict its pK_a values with CpHMD. The presence of a small pK_a profile for Asp25 strengthens the idea that pHRE is a good method to improve the sampling in systems like pHLIP.

We also observe a striking pattern between the different profiles, where the global shift of the residues is similar to the shift observed for Asp14 in the other simulations, corroborating the desolvation effect. However, the shifts of these residues are not only different from the observed for Asp14, but the magnitude of the shift is correlated with their sequence position. In the water phase, all residues start from regular pK_a values and their differences are more or less correlated with their bulk pK_a values. Along the membrane normal, each residue pK_a will shift in the same direction but with different magnitudes. Using Asp14 as a reference, the pK_a shift of Asp31 is larger by 1 pK_a unit. Since the pK_a of Asp14 only depends on the desolvation effect, this means that there is electrostatic repulsion, from its neighbor acidic residues, taking place. Indeed, when inserting, Asp31 will be protonating before any of its neighbors (Asp33, Glu34, C-ter), since these will not be in such inserted positions. The neighbor residues negative charges will have a significant impact on the pK_a values of Asp31, increasing its shift to higher values. When Asp33 reaches this region, Asp31 is long protonated leaving one less negative charge in the neighborhood, hence resulting in a smaller pK_a shift for this residue.

These results suggest that Asp31 will be the first to fully insert in the withdrawal process assuming that Asp25 is almost always protonated in pHLIP state III. Following this logic, the next residue to insert should be Asp33, since it bears the second highest pK_a shift of the C-terminus residues. This phenomenon was already observed in the CpHMD results, but we lacked information for several other residues. Fundamentally, we find evidence that the observed order is correlated with a pK_a shift dependent withdrawal process.

3.3 L16H pHLIP Simulations

In the previous subsection, we concluded that, for the pHLIP system, the best setup would be: a membrane of 128 POPC lipids, a pH range of 4.0 to 7.5 with 0.5 pH steps and a τ_{RE} of 20 ps. However, we also extended the simulations of the new system to the use of a τ_{RE} of 100 ps. At the present time, the simulations for the system with τ_{RE} 20 ps are still running and only the τ_{RE} 100 can be presented.

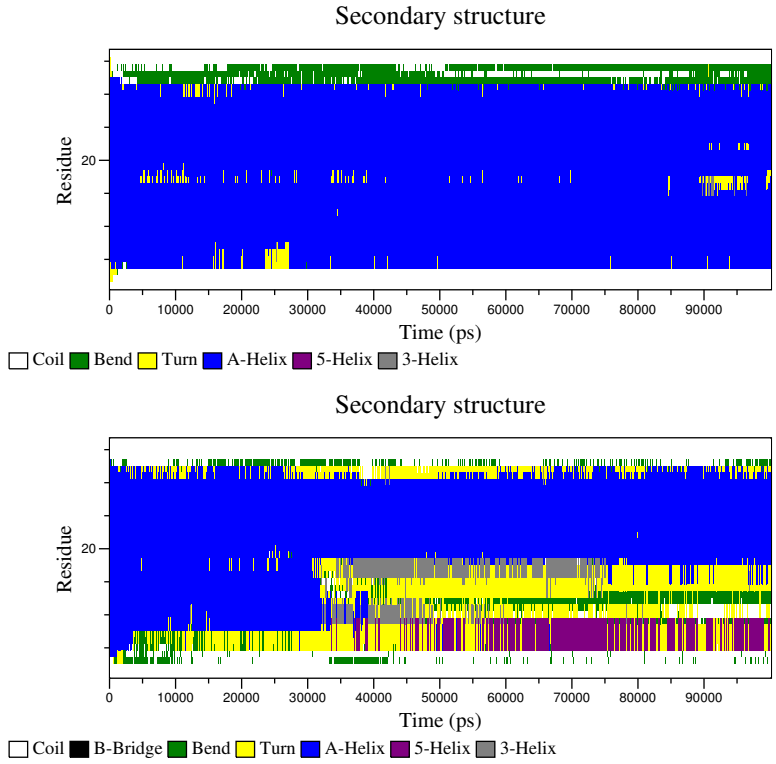

Figure 3.11: Secondary structure representation (DSSP criteria) of two CpHMD replicates of L16H pHLIP peptide at pH 4.0 (replicate 2 in **top** and replicate 3 in **bottom**).

Figure 3.11 illustrates two different replicates where the L16H pHLIP system can sample two quite different conformational regions. One where the helix integrity is not affected (Figure 3.11 **top**), similar to what is more commonly observed in *wt*-pHLIP, and another one where the N-terminal region of the α-helix is destabilized (Figure 3.11 **bottom**), probably induced by the need to solvate the abundantly protonated His16 residue.

Similarly to the *wt*-pHLIP analysis, the percentage of helical content in the pHRE simulations is consistent with the expected conformations in state III. The results are also in agreement with the CpHMD results as evidenced by Figure 3.12. Therefore, we should be sampling the state III of pHLIP and all predicted pK_a values and be compared with *wt*-pHLIP. Note that the presence of the histidine at 16^{th} position could have an effect on Asp14, which would lead to lower pK_a values. However, this does not happen and the pK_a values observed are in the same region (Figure 3.13). This result is probably a consequence of our decision to insert the His residue at the 16^{th} position. This was chosen because, in an α-helix, Asp14 and His16 are facing opposite sides of pHLIP. There is a large distance between both residues and since the electrostatic interactions are distance dependent, the effect of His16 on Asp14 will be small.

A good description of Asp14 protonation profile is of major importance due to its trigger role

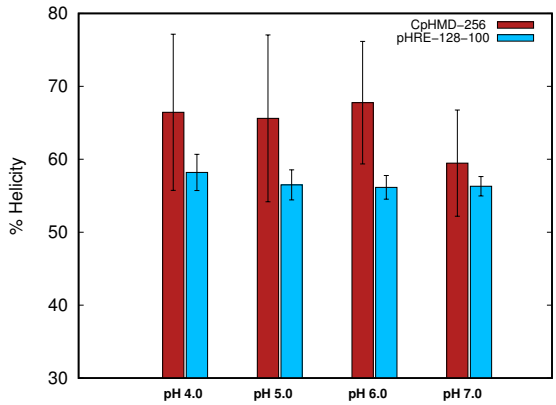

Figure 3.12: Percentage of helical content for the L16H variant of pHLIP. Comparison between the CpHMD (256 lipids) and pHRE simulations (128 lipids and τ_{RE} 100ps).

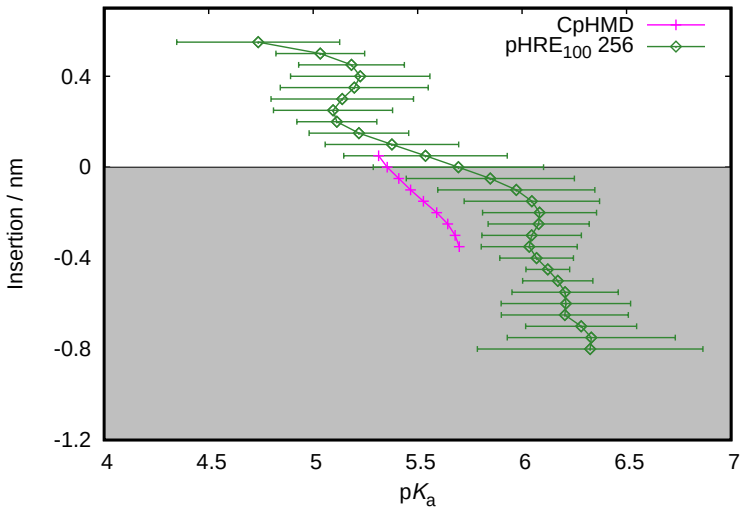

Figure 3.13: pK_a profile of Asp14 obtained from CpHMD and pHRE simulations (τ_{RE} 100 ps) of the L16H variant of pHLIP. The CpHMD results do not have sufficient data to calculate error bars, but the points were plotted anyway for comparison.

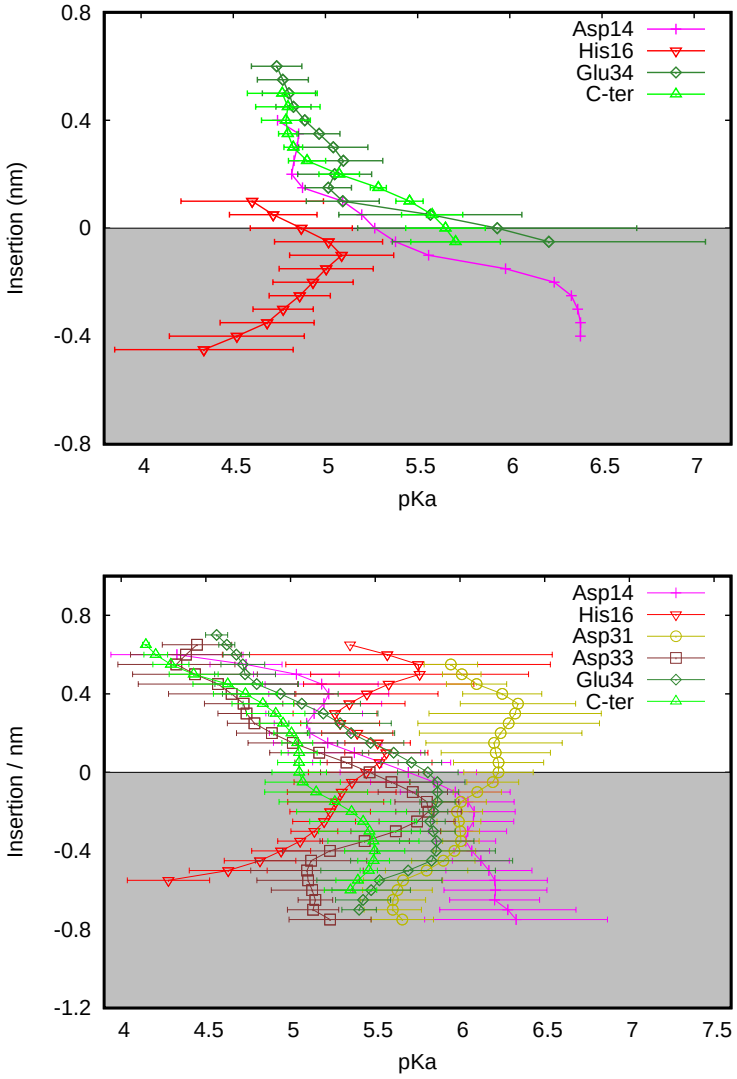

Figure 3.14: pK_a profiles for all the titrable residues in the L16H variant of pHLIP using CpHMD (**top**) or pHRE with τ_{RE} 100 ps (**bottom**). All the profiles have fulfilled the required criteria with the exception of Asp14 in CpHMD simulations, due to lack of sampling.

in the insertion process of pHLIP. Predicting a pK_a profile for Asp14 also requires a good error estimation, otherwise our results do not possess statistical significance. Therefore, the inability to calculate errors for the CpHMD profile of Asp14 speaks volumes of the limitations that CpHMD has in sampling this system. As it stands, the pHRE offers a much better description of the Asp14 pK_a profile then the CpHMD.

The pK_a profile can provide good insights on the core interactions of the peptide (Figure 3.14). Again, the residues outside the membrane start with a small pK_a that shifts towards higher pK_a values except for Asp31 in pHRE simulations (Figure 3.14 **bottom**). This residue presents an interesting profile since its initial pK_a (∼5.7) away from the membrane is already significantly high. This means that Asp31 is the initially protonated residue, even in solution. The high value in the pK_a can probably be explained by the strong electrostatic repulsion from its neighbors (Asp33, Glu34 and C-ter). The presence of negative electrostatic potential increases the initial pK_a towards 6.0 until 6.3 at 0.4 nm from the membrane surface. Notably, the pK_a values start to decrease

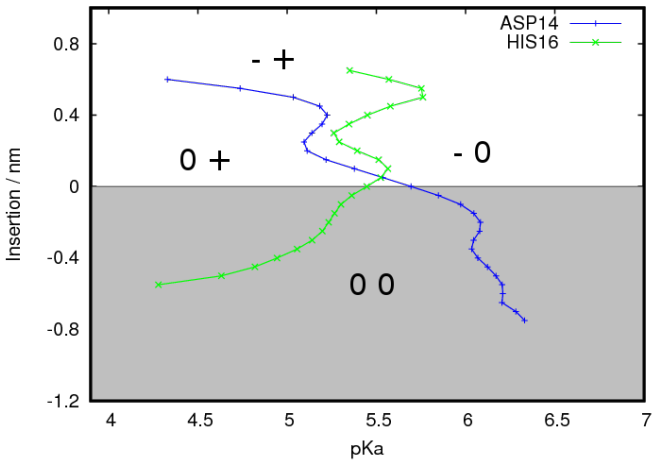

Figure 3.15: pK_a profile of Asp14 and His16 obtained from pHRE simulations of the L16H variant of pHLIP. The most probable charges are indicated in each region of the plot.

from that point on, in an opposite note from the other anionic residues observed until now. Take into consideration that, at 0.4 nm for Asp31, the other residues are a bit farther away from the membrane only feeling the electrostatic potential. However, with insertion they start to feel the low dielectric of the membrane and their pK_a values will further increase. The higher the pK_a values from those residues become, the probability of them protonating increases. As more events of protonated states for those residues appear, the magnitude of the negative electrostatic potential decreases. Thus, the pK_a of Asp31 will diminish if the reduction of the electrostatic repulsion overpowers the desolvation effect. This phenomena can probably be observed for every residue in the C-terminus domain, even in the *wt*-pHLIP, yet not as evident. The Asp33 residue starts to experience this effect near -0.2 nm of insertion (Figure 3.14 **bottom**), as the pK_a starts to shift towards lower values, we expect Glu34 and C-ter to be slightly inserted or at the membrane surface, hence a higher pK_a and more probable protonated states. At the deepest insertion, we observe that the pK_a shifts for those residues begins to increase because, at this point, all the residues are most likely protonated and the desolvation effect will prevail, increasing the pK_a values.

These pK_a profiles define the complexity behind the pHLIP system and the importance of robust and accurate methodologies to sample protonation states and define membrane insertions. Hence, the CpHMD method falls short in defining the interactions of this system and characterizing some residues, like the unreliable pK_a predictions on Asp14 whose profile is merely representative without any statistical significance due to the lack of sampling.

One of our main objectives was to introduce a cationic residue that would provide an additional pK_a^{ins} thereby delimiting the pH range of insertion. The rationale follows that a cationic residue at lower pH values would ionize and force the peptide to exit the membrane. This charge balance between the anionic and cationic regulators of pHLIP dictates the inserted stability of the peptide. Figure 3.15 shows the pK_a profile of the two possible regulators of pHLIP insertion: Asp14 (pK_a 6.3) and His16 (pK_a 4.2). Outside the membrane, we observe that the two residues are ionized most of the time. As they approach the membrane, the low dielectric starts to affect both residues, leading to a shift in the pK_a values favoring their neutral species. While, for an anionic residue, their pK_a shifts towards higher values, increasing the probability of being protonated (neutral), the pK_a value of a cationic residue starts to decrease as it gets closer to the membrane. The apolar

environment of the membrane and the desolvation effect forces the residues to stay neutral, thus the His16 pK_a shifts, increasing the probability of deprotonation (neutral). We also observe a mid-point where the two profiles cross over each other, near the surface of the membrane. This mid-point is important because it describes the precise insertion region where the probability of the residues to be neutral becomes larger than the probability of them being ionized. The region below that mid-point is characterized by a $\Delta pK_a^{ins}(pK_a^{ins}{}_{Asp} - pK_a^{ins}{}_{His})$. Two experimental p$K_{ins}$ (3.3 and 5.9) are known for the L16H variant which were measured by your collaborators [40].

Note that the higher pK_a^{ins} is almost equal to the pK_{ins} of *wt*-pHLIP, indicating that the presence of the histidine does not influence the insertion process dependent of Asp14. This is important since this objective was in our initial design of the His mutation. The experimental ΔpK_{ins} is approximately 2.6 pK_a units, suggesting that pHLIP may insert in a large range of pH values. The ΔpK_a^{ins} measured for both the LRA method (2.75 pK_a units) and the pHRE method (2.1 pK_a units) is in agreement with the experimental result within the errors of the methods. This result also emphasizes our capability to describe pHLIP with pHRE and predict not only the pK_a^{ins} values of key residues, but also the range of pH values where pHLIP inserts. However, the aforementioned results suggest that the low pK_a^{ins} value of His16 (4.3 for pHRE and 3.75 for LRA) might induce a too large ΔpK_a^{ins} to be useful for *in vivo* studies and clinical purposes. Nevertheless, the results provide a good background for new pHLIP sequences to be studied in order to narrow down the pK_a^{ins} range to one pK_a unit or even below. In the end, we want to be able to discriminate between tumoral cells and between these and normal cells whose pH is acidic.

Chapter 4

Ongoing Work

Throughout this thesis, we presented the development of important groundwork for the future. Not only a new insertion method came to fruition, but different methodologies were used to study the pHLIP system: the CpHMD and pHRE methods. The *wt*-pHLIP and L16H variant were extensively studied using the two methodologies, providing interesting results regarding the pK_a^{ins} values of key residues. As such, the study of more variant sequences of *wt*-pHLIP with the pHRE methodology is the next logical step. Some of these sequences were simulated in previous works using the LRA methodology [40] where the position of Asp14 was changed to the 13^{th} or 15^{th} position. These first few tests will provide a calibrating measure of the method and, possibly, obtain more evidences of the correlation between the experimental pK_{ins} and the pK_a^{ins} of the Asp14 residue. Additionally, we will test other sequences [9] suggested by Professor Andreev and evaluate the changes introduced by other point mutations on pHLIP in state III. Another future perspective is the improvement of the ΔpK_a^{ins} observed with the L16H variant. In a first instance, by changing the position of the His residue, the electrostatic interactions with Asp14 may change and affect the conformational space of the peptide. Hence, we expect the ΔpK_a^{ins} to be reduced and, moreover, assess how the position of the cationic residue influences the insertion process. With this knowledge, we think different cationic residues will change the ΔpK_a^{ins} and further fine tune the insertion pH range. We will test lysine and its non-natural shorter side-chain derivatives (ornithine, diaminobutyric acid and diaminopropionic acid). A shorter side-chain prohibits an ionic interaction between the cationic and anionic residues, thus favoring the neutral states of the residue instead of the ionized states, allowing the insertion of the peptide. Overall, there are a lot of possibilities to be explored and further improve both our understanding and the therapeutic applications of this complex system.

Chapter 5

Concluding Remarks

In this thesis, the conformational changes induced by pH effects on *wt*-pHLIP and its L16H variant were our major focus. For this task, we devised three techniques to study the effect of insertion on the environment of the residues and how each method influences key residues pK_a predictions. The used insertion method is crucial to define the gradient of solvent exposure for each residue and an unreliable description results in a poor pK_a profile. The new distance cutoff method showed to be accurate and robust in describing the gradient of solvent exposure for a given residue, thus improving our pK_a estimations.

By studying the residues pK_a profiles, we may deduce how likely a given residue is inserted at any given pH value and how these residues influence each other electrostatically, like in the C-terminus domain. These informations shed light upon the insertion/withdrawal mechanism of pHLIP at the molecular level, pushing us towards new pHLIP sequences with fine tuned features. Although the L16H variant failed to provide a narrow pH range of insertion, it was a good proof of concept that is possible to stabilize the state III in the pK_a^{ins} range between Asp14 and His16. This concept was first explored using LRA simulations and, in this work, it was further extended with CpHMD and, finally, with the pHRE method. Although the LRA and CpHMD methods were able to predict pK_a values to some extent, they presented some concerning limitations in sampling.

The pHRE method was under scrutiny during a large part of this work, in order to optimize simulations and, hopefully, overcome sampling limitations of the other methods. We simulated several systems to compare with the CpHMD method and we concluded that pHRE is able to sample and predict pK_a^{ins} values in good agreement with the experimental results. Overall, the pHRE simulations show remarkable results by surpassing the limitations of the LRA and CpHMD methodologies, achieving better sampling in deeper membrane regions, where the pK_a values are higher, corroborating the expected solvent effect.

The pH-dependent conformational changes of pHLIP were extensively studied using different simulation groups. In *wt*-pHLIP, we observed a strong correlation between the experimental pK_{ins} and the pK_a of Asp14 at the deepest measured insertion (pK_a^{ins}). The insertion process seems to be dependent on the protonation state of this residue, acting as a trigger. The residues in the C-terminus region seems to insert in an ordered sequence determined by their pK_a shifts. Thus, the insertion and withdrawal processes of pHLIP are defined by the charge balance in the C-terminus and the protonation state of Asp14.

The peptide remained in state III after the addition of an histidine at the 16^{th} position. We observed that the L16H pHLIP only inserts in the membrane if both residues are neutral, successfully delimiting a range of insertion. The key location of the cationic residue, in an opposite position of

Asp14, leads to an almost unnafected pK_a^{ins} for this acid, which was one of the intended goals of the variant design. However, the pK_a^{ins} of His16 was too low, giving rise to a too large ΔpK_a at which the peptide should remain inserted. To properly achieve a therapeutic application, this pH range of insertion needs to be more narrow and specific. We propose that residues with higher pK_a values will be more successful in closing down the range of insertion, thus improving clinical application.

In sum, the pHRE method has proven to be an important step forward in studying not only the pHLIP system but other complex systems, where conformational and protonation sampling is limited. Despite this thesis important contribution to the understanding of the pHLIP insertion and withdrawal processes, there are still many question open. In the future, several new sequences will be studied with this methodology and, hopefully, improve the therapeutic capabilities of pHLIP technologies.

Bibliography

[1] D Lee Nelson and Michael M Cox. *Lehninger Principles of Biochemistry*. New York, New York: WH Freeman and Company, 2005.

[2] Bruce Alberts. *Molecular biology of the cell*. Garland science, 2002.

[3] Florian Cymer, Gunnar von Heijne, and Stephen H White. Mechanisms of integral membrane protein insertion and folding. *Journal of molecular biology*, 427(5):999–1022, 2015.

[4] Y. K. Reshetnyak, M. Segala, O. A. Andreev, and D. M. Engelman. A monomeric membrane peptide that lives in three worlds: in solution, attached to, and inserted across lipid bilayers. *Biophys. J.*, 93(7):2363–2372, 2007.

[5] Oleg A Andreev, Allison D Dupuy, Michael Segala, Srikanth Sandugu, David A Serra, Clinton O Chichester, Donald M Engelman, and Yana K Reshetnyak. Mechanism and uses of a membrane peptide that targets tumors and other acidic tissues in vivo. *Proceedings of the National Academy of Sciences*, 104(19):7893–7898, 2007.

[6] Monika Musial-Siwek, Alexander Karabadzhak, Oleg A Andreev, Yana K Reshetnyak, and Donald M Engelman. Tuning the insertion properties of pHLIP. *BBA-Biomembranes*, 1798(6):1041–1046, 2010.

[7] Oleg A Andreev, Alexander G Karabadzhak, Dhammika Weerakkody, Gregory O Andreev, Donald M Engelman, and Yana K Reshetnyak. ph (low) insertion peptide (phlip) inserts across a lipid bilayer as a helix and exits by a different path. *Proceedings of the National Academy of Sciences*, 107(9):4081–4086, 2010.

[8] O. A. Andreev, D. M. Engelman, and Y. K. Reshetnyak. Targeting diseased tissues by pHLIP insertion at low cell surface pH. *Front. Physiol.*, 5, 2014.

[9] Dhammika Weerakkody, Anna Moshnikova, Mak S Thakur, Valentina Moshnikova, Jennifer Daniels, Donald M Engelman, Oleg A Andreev, and Yana K Reshetnyak. Family of pH (low) insertion peptides for tumor targeting. *Proc. Natl. Acad. Sci. USA*, 110(15):5834–5839, 2013.

[10] Matthew G Vander Heiden, Lewis C Cantley, and Craig B Thompson. Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033, 2009.

[11] Marion Stubbs, Paul MJ McSheehy, John R Griffiths, and C Lindsay Bashford. Causes and consequences of tumour acidity and implications for treatment. *Molecular medicine today*, 6(1):15–19, 2000.

[12] Morteza Naghavi, Reji John, Sameh Naguib, Mir Said Siadaty, Roxana Grasu, KC Kurian, W Barry van Winkle, Babs Soller, Silvio Litovsky, Mohammad Madjid, et al. ph heterogeneity of human and rabbit atherosclerotic plaques; a new insight into detection of vulnerable plaque. *Atherosclerosis*, 164(1):27–35, 2002.

[13] Justin Fendos, Francisco N Barrera, and Donald M Engelman. Aspartate embedding depth affects pHLIP's insertion pK_a. *Biochemistry*, 52(27):4595–4604, 2013.

[14] Francisco N Barrera, Dhammika Weerakkody, Michael Anderson, Oleg A Andreev, Yana K Reshetnyak, and Donald M Engelman. Roles of carboxyl groups in the transmembrane insertion of peptides. *Journal of molecular biology*, 413(2):359–371, 2011.

[15] Wilfred F van Gunsteren and Herman JC Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition*, 29(9):992–1023, 1990.

[16] Wilfred F van Gunsteren, Dirk Bakowies, Riccardo Baron, Indira Chandrasekhar, Markus Christen, Xavier Daura, Peter Gee, Daan P Geerke, Alice Glättli, Philippe H Hünenberger, et al. Biomolecular modeling: goals, problems, perspectives. *Angewandte Chemie International Edition*, 45(25):4064–4092, 2006.

[17] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[18] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, and David E Shaw. Picosecond to millisecond structural dynamics in human ubiquitin. *J. Phys. Chem. B*, 120(33):8313–8320, 2016.

[19] A. R. Leach. *Molecular modelling: principles and applications*. Addison-Wesley Longman Ltd, 2001.

[20] M P Allen and D J Tildesley. *Computer Simulation of Liquids*. Oxford University Press, USA, 1987.

[21] Menahem Pirchi, Guy Ziv, Inbal Riven, Sharona Sedghani Cohen, Nir Zohar, Yoav Barak, and Gilad Haran. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nature communications*, 2:493, 2011.

[22] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[23] Miguel Machuqueiro and António M. Baptista. The pH-dependent conformational states of kyotorphin: A constant-pH molecular dynamics study. *Biochem. J.*, 92:1836–1845, 2007.

[24] António M. Baptista and Cláudio M. Soares. Some theoretical and computational aspects of the inclusion of proton isomerism in the protonation equilibrium of proteins. *J. Phys. Chem. B*, 105:293–309, 2001.

[25] António M. Baptista, Vitor H. Teixeira, and Cláudio M. Soares. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.*, 117:4184–4200, 2002.

[26] Miguel Machuqueiro and António M. Baptista. Constant-pH molecular dynamics with ionic strength effects: Protonation–conformation coupling in decalysine. *J. Phys. Chem. B*, 110:2927–2933, 2006.

[27] Miguel Machuqueiro and António M. Baptista. Acidic range titration of HEWL using a constant-pH molecular dynamics method. *Proteins Struct. Funct. Bioinf.*, 72:289–298, 2008.

[28] Miguel Machuqueiro and António M. Baptista. Molecular dynamics constant-pH and reduction potential: Application to cytochrome c_3. *J. Am. Chem. Soc.*, 131:12586–12594, 2009.

[29] Miguel Machuqueiro, Sara RR Campos, Cláudio M Soares, and António M Baptista. Membrane-induced conformational changes of kyotorphin revealed by molecular dynamics simulations. *J. Phys. Chem. B*, 114(35):11659–11667, 2010.

[30] Vitor H Teixeira, Diogo Vila-Viçosa, António M Baptista, and Miguel Machuqueiro. Protonation of dmpc in a bilayer environment using a linear response approximation. *J. Chem. Theory Comput.*, 10:2176–2184, 2014.

[31] Hugo AF Santos, Diogo Vila-Viçosa, Vitor H Teixeira, António M Baptista, and Miguel Machuqueiro. Constant-pH MD simulations of DMPA/DMPC lipid bilayers. *J. Chem. Theory Comput.*, 11(12):5973–5979, 2015.

[32] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1):141–151, 1999.

[33] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[34] Jae-Yel Yi, Jerry Bernholc, and Peter Salamon. Simulated annealing strategies for molecular dynamics. *Computer physics communications*, 66(2-3):177–180, 1991.

[35] Jason A Wallace and Jana K Shen. Continuous constant pH molecular dynamics in explicit solvent with ph-based replica exchange. *Journal of chemical theory and computation*, 7(8):2617–2629, 2011.

[36] B. H. Morrow, P. H. Koenig, and J. K. Shen. Self-assembly and bilayer–micelle transition of fatty acids studied by replica-exchange constant pH molecular dynamics. *Langmuir*, 29(48):14823–14830, 2013.

[37] Satoru G Itoh, Ana Damjanović, and Bernard R Brooks. ph replica-exchange method based on discrete protonation states. *Proteins Struct. Funct. Bioinf.*, 79(12):3420–3436, 2011.

[38] Jason M Swails, Darrin M York, and Adrian E Roitberg. Constant ph replica exchange molecular dynamics in explicit solvent using discrete protonation states: implementation, testing, and validation. *J. Chem. Theory Comput.*, 10(3):1341–1352, 2014.

[39] D. Vila-Viçosa, A. M. Baptista, C. Oostenbrink, and M. Machuqueiro. A pH replica exchange scheme in the stochastic titration constant-pH MD method. *In preparation*.

[40] D. Vila-Viçosa, Oleg A. Andreev, , and M. Machuqueiro. pKa values of pHLIP key amino acids at the water / membrane interface. *In preparation*.

[41] Ilario G. Tironi, René Sperb, Paul E. Smith, and Wilfred F. van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.*, 102:5451–5459, 1995.

[42] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:014101, 2007.

[43] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, dec 1981.

[44] DVD Spoel, E Lindahl, B Hess, ARVan Buuren, E Apol, PJ Meulenhoff, DP Tieleman, ALTM Sijbers, KA Feenstra, RVan Drunen, et al. Gromacs user manual version 4.0. *Gromacs, Groningen, The Netherlands*, 2009.

[45] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[46] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, 1995.

[47] P Beroza, D R Fredkin, M Y Okamura, and G Feher. Protonation of interacting residues in a protein by a monte carlo method: application to lysozyme and the photosynthetic reaction center of rhodobacter sphaeroides. *Proc. Natl. Acad. Sci. USA*, 88(13):5804–5808, 1991.

[48] S. R. R. Campos, M. Machuqueiro, and A. M. Baptista. Constant-ph molecular dynamics simulations reveal a β-rich form of the human prion protein at low ph. *J. Phys. Chem. B*, 114:12692–12700, 2010.

[49] Luís CS Filipe, Sara RR Campos, Miguel Machuqueiro, Tamis Darbre, and António M Baptista. Structuring peptide dendrimers through ph modulation and substrate binding. *J. Phys. Chem. B*, 120(38):10138–10152, 2016.

[50] N. Schmid, A.P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A.E. Mark, and W.F. Van Gunsteren. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.*, 40(7):843–856, 2011.

[51] B. Hess. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.*, 4(1):116–122, 2008.

[52] Jan Hermans, Herman J. C. Berendsen, Wilfred F. van Gunsteren, and Johan P. M. Postma. A Consistent Empirical Potential for Water-Protein Interactions. *Biopolymers*, 23:1513–1518, 1984.

[53] P.E. Smith and W.F. van Gunsteren. Consistent dielectric properties of the simple point charge and extended point charge water models at 277 and 300 K. *J. Chem. Phys.*, 100:3169–3174, 1994.

[54] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.*, 23(1):128–137, 2002.

[55] L. Li, C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, L. Wang, N. Smith, M. Petukh, and E. Alexov. DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys.*, 5(1):9, 2012.

[56] Vitor H. Teixeira, Carlos C. Cunha, Miguel Machuqueiro, A. Sofia F. Oliveira, Bruno L. Victor, Cláudio M. Soares, and António M. Baptista. On the use of different dielectric constants for computing individual and pairwise terms in Poisson-Boltzmann studies of protein ionization equilibrium. *J. Phys. Chem. B*, 109:14691–14706, 2005.

[57] Frederic M Richards. Areas, volumes, packing, and protein structure. *Annual review of biophysics and bioengineering*, 6(1):151–176, 1977.

[58] M.K. Gilson, K.A. Sharp, and B. Honig. Calculating the eletrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.*, 9:327–335, 1987.

[59] António M. Baptista, Paulo J. Martel, and Cláudio M. Soares. Simulation of electron-proton coupling with a Monte Carlo method: Application to cytochrome c_3 using continuum electrostatics. *Biophys. J.*, 76:2978–2998, 1999.

[60] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[61] Vitor H Teixeira, Cristina Ventura, Ruben Leitão, Clara Ràfols, Elisabeth Bosch, Filomena Martins, and Miguel Machuqueiro. Molecular details of INH–C10 binding to wt katG and its S315T mutant. *Mol. Pharm.*, 12(3):898–909, 2015.

[62] A. McIntosh. The Jackknife Estimation Method. *ArXiv e-prints*, June 2016.

[63] Catarina A. Carvalheda, Sara R. R. Campos, and António M. Baptista. The effect of membrane environment on surfactant protein C stability studied by Constant-pH molecular dynamics. *J. Chem. Inf. Model.*, 55(10):2206–2217, 2015.

[64] Norbert Kučerka, Mu-Ping Nieh, and John Katsaras. Fluid phase lipid areas and bilayer thicknesses of commonly used phosphatidylcholines as a function of temperature. *Biochem. Biophys. Acta, Biomembr.*, 1808:2761–2771, 2011.

[65] Vitor H Teixeira, Ana Sofia C Capacho, and Miguel Machuqueiro. The role of electrostatics in trxr electron transfer mechanism: A computational approach. 84(12):1836–1843, 2016.